

# GEODESY: THE CONCEPTS

Petr VANÍČEK  
Edward J. KRAKIWSKY\*  
*University of New Brunswick  
Canada*

(\*Now at the University of Calgary)

Second Edition



1986  
NORTH-HOLLAND  
AMSTERDAM · NEW YORK · OXFORD · TOKYO

© Elsevier Science Publishers B.V., 1986

*All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.*

First edition 1982  
Second edition 1986

ISBN: 0444 87775 4

*Publishers*  
ELSEVIER SCIENCE PUBLISHERS B.V.  
P.O. BOX 1991, 1000 BZ AMSTERDAM  
THE NETHERLANDS

*Sole distributors for the U.S.A. and Canada*  
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.  
52 VANDERBILT AVENUE  
NEW YORK, NY 10017  
U.S.A.

**Library of Congress Cataloging in Publication Data**

Vaniček, Petr, 1935-  
Geodesy, the concepts.

Bibliography: p.  
Includes index.  
I. Geodesy. I. Krakiwsky, Edward J., 1938-  
II. Title.  
QB281.V36 1985 526'.1 85-10156  
ISBN 0-444-87775-4  
ISBN 0-444-87777-0 (pbk.)

PRINTED IN THE NETHERLANDS

*To Jana and Myrna for their understanding,  
and to all our friends  
for their encouragement*

## FOREWORD

For many years we have felt that there was a definite need for a new comprehensive textbook on geodesy that would (a) deal with the totality of geodesy, (b) treat it in a conceptual manner, and (c) integrate the recent developments with the standard geodetic topics. Our experience has led us to believe that the new ideas and techniques introduced into geodesy in the past three or four decades have so significantly changed the character of geodesy as to require a new delimitation of its scope. This we have attempted to do in §4.1. As a by-product of this delimitation came the *functionalization* of geodesy, which is reflected in the structure of the book. The concepts needed for the three major functions of geodesy are described in the last three parts: positioning (IV), the study of the earth's gravity field (V), and the study of the earth's temporal deformations (VI).

A secondary goal of this book has been the *demystification* of geodesy. To this end we have tried to clarify the terminology and make it as uniform as possible. There are no more physical and geometrical geodesies; there is no satellite geodesy, no vertical geodesy, and no kinematic geodesy. Wherever appropriate, we have also attempted to synthesize and classify the ideas involved. When terms from other fields were needed, we have conscientiously tried to use them in their original form thus we use, for example, tide gauge instead of medimaremeter, confidence region rather than error ellipse. Whenever possible, however, we have used the prevailing terminology as well as the existing symbols.

The choice of concentrating on *concepts* was made for three reasons: (a) to make the book as helpful as possible for the student who wants to learn geodesy on his own, (b) to keep the book brief, and (c) to prevent the contents from aging too quickly. The price we had to pay for this decision was that many things had to be omitted. For instance, most of the proofs had to be left out so that the reader, as well as the instructor who might use the book for his course, is required to fill in the gaps. This we considered preferable, however, to the alternative of spelling out the proofs and letting the reader distill the concepts from the maze of techniques. We have tried to make the task of bridging the gaps easier by pointing out the assumptions involved, the major steps in the reasoning chain, and the appropriate mathematical apparatus needed. Many new, and as yet untested, ideas fell victim to our policy of brevity. Also many side issues and less important concepts could only be given a reference to the literature.

No attempt has been made to include descriptions of measuring techniques and instrumentation. We believe that those are more properly treated in a textbook on surveying. Thus only the concepts of the necessary surveying techniques are shown to facilitate the understanding of the nature of the different kinds of data collected in the field. On the other hand, we have included a whole part (III) that deals with the mathematical *methodology of geodesy*. By having gathered in one place all the needed mathematical techniques, we believe we have saved space and the reader's time.

The book contains material suitable for either technological courses or university undergraduate and graduate courses. For instance, Parts I and II could constitute an introductory geodesy course either on the technological or university levels. For technological schools, such a course could be complemented by a course assembled from selected material from Part IV. For a university, the follow-up courses could consist of Part IV and selected material from Parts V and VI. At the other end of the spectrum, Part VI would probably be considered by most as being predominantly of a graduate nature. These questions are discussed in some detail in §4.4.

Certain features of this book, designed to assist the student and instructor alike, should be mentioned here. First, each chapter has an unnumbered *introductory paragraph* meant to give an overview of the material contained in that chapter. In addition, these overviews sometimes contain concepts which are not spelled out anywhere else. The most *important formulae* are boxed in; these formulae are generally referred to more often than the others. Similarly, the *key terms*, when introduced for the first time, are in italics; they are listed in the Subject Index, where the number refers to the page on which the term is defined. With "a good picture being worth a thousand words", much thought and work has gone into the *illustrations*, and as many concepts as possible have been shown graphically. We have also done our best to select appropriate *references* following the rule that they should be available in English and accessible in the open literature. There are, however, a few inevitable exceptions to this rule. References are gathered at the end of each part; hence, some publications are listed in several places. Equations, figures, and tables are *numbered* separately in each chapter. When a reference is made to them within the same chapter, the chapter number is omitted. If an equation, figure, or table outside the chapter is referred to, its number appears in full.

The fourteen sets of lecture notes we have written for various courses offered by the Department of Surveying Engineering of the University of New Brunswick have been the basis for this book. The material presented herein has thus been tested through teaching on both the undergraduate and graduate levels. Innumerable discussions carried out with our colleagues as well as students over the years have helped us to form some of the opinions and reach some of the insights presented here. The emphasis has been placed on clarity rather than originality. This should be clear to the reader from the many references cited in the book by which we have tried to disclaim parenthood for the vast majority of ideas presented here.

Many people contributed to our effort by commenting or critiquing different portions of the manuscript, and by diverse personal communications. We hereby gratefully acknowledge the assistance of Mr. J.R. Adams, Prof. E.G. Anderson, Prof.

J A R. Blais, Dr. G. Blaha, Prof. C. Beaumont, Dr. J.D. Bossler, Mr. W.H. Falkenberg, Dr. K. Frankich, Prof. C. Gemael, Prof. E.W. Grafarend, Mr. L.F. Gregerson, Dr. B. Guinot, Prof. A.C. Hamilton, Prof. F. Hatschbach, Dr. R.C. Jachens, Prof. W.R. Knight, Mr. J. Kouba, Mr. M.P. Mepham, Dr. D. Nagy, Prof. N. Ni Chuiv, Mr. B.G. Nickerson, Dr. M.K. Paul, Mr. A.J. Pope, Prof. S. Rinco, Prof. M.G. Rochester, Prof. K-P. Schwarz, Dr. R.A. Snay, Mr. R.R. Steeves, Prof. J.H. Thompson, Dr. D.B. Thomson, Prof. R.S. Turner, Prof. D.E. Wells, Dr. C.A. Whitten, Mr. T. Wray, and Dr. S. Yumi. Much of the credit for the book is due to these people; any blame, however, is solely ours for, perhaps, not always heeding their advice. We shall be grateful for any constructive criticism communicated to us in the future.

In conclusion, we should like to express our special thanks to Ms. Wendlynn Wells for her editing of the manuscript into readable English, her faultless typing of the many versions of the manuscript, and the constant pressure she exerted on us. Without this pressure, the book would probably never have seen the light of day. Credit for the illustrations goes to Mr. M. Anderson, Mrs. V. Rinco and Mrs. D. Jordan. Many more people have been instrumental in the preparation of the manuscript including our colleagues in the Department who have had to shoulder some of our academic responsibilities while we wrote the book. To all these we are truly grateful.

*Petr Vaníček  
Edward J. Krakiwsky*

*Fredericton, N.B., Canada  
26 September 1980*

## **FOREWORD TO THE SECOND EDITION**

It is with great pleasure that we are adding to the original Foreword on the occasion of the second edition of this book. The first edition was sold out within a relatively short time. We take this to be proof that there was a need for a comprehensive textbook on geodesy, and thus our many years of labour seem to have been justified. The new edition gives us the opportunity to correct the most glaring errors that wiggled their way into the first printing.

After reading the published reviews we felt that there probably was no reason to restructure the book, the philosophy of our approach having been received generally in a very positive vein. Consequently, there was no need to change the Foreword to the first edition, which still applies as much as it did originally. We have instead concentrated on updating the book by adding the Geodetic Reference System 1980; a fuller treatment of the Global Positioning System (NAVSTAR), and of satellite altimetry. As well, results of some recent investigations of sea surface topography, of strain in geodetic networks, of topographic correction to gravity, of the systematic effect of geodetic reference ellipsoid misalignment on different kinds of geodetic quantities were introduced. The geodetic boundary value problem has been reformulated, and the vertical gradient of gravity has been restated. Many bumpy formulations were smoothed, several new important references added. All this was done under a self-imposed, severe space constraint.

Many people communicated to us their findings of mistakes and misleading statements. To all of these we are truly thankful. Special thanks, however, go to Dr. I. Bauersima, Dr. A. Kleusberg, Prof. R.B. Langley, and Dr. I. Reilly for their thoughtful comments. As usual, graduate students were very helpful by critiquing the book: Messrs. G. Carrera, N. Christou, M. Craymer, and P. Tetreault in particular deserve mention here. As with the first edition, we have relied heavily on the editorial skills and devotion of Ms. Wendy Wells; for both we are truly thankful.

In spite of the valiant effort of so many of us, some errors and mistakes will have inevitably survived the hunt. We will thus continue appreciating any constructive criticism communicated to us.

Because in the past several years we have been physically separated by some 3500 km, close cooperation on the new edition was virtually impossible. The senior author alone, therefore, should be blamed for any additional errors he may have unwittingly introduced when carrying out the changes.

*Fredericton, N.B., Canada  
1 October 1984*

*Petr Vaníček  
Edward J. Krakiwsky*

## **CONTENTS**

<b>FOREWORD</b>	vii
<b>FOREWORD TO THE SECOND EDITION</b>	x
<b>PART I. INTRODUCTION</b>	1
<b>    1. HISTORY OF GEODESY</b>	3
1.1. Historical beginnings of geodesy	4
1.2. Scientific beginnings of geodesy	8
1.3. Geodesy in the service of mapping	14
1.4. Geodesy of the modern era	16
<b>    2. GEODESY AND OTHER DISCIPLINES</b>	19
2.1. Applications of geodesy	19
2.2. Symbiotic relation between geodesy and some other sciences	21
2.3. Theoretical basis of geodesy	23
<b>    3. MATHEMATICS AND GEODESY</b>	25
3.1. Algebra	25
3.2. Analysis	30
3.3. Geometry	37
3.4. Statistics	41
<b>    4. STRUCTURE OF GEODESY</b>	45
4.1. Functions of geodesy	45
4.2. Geodetic theory	47
4.3. Geodetic practice	48
4.4. Geodetic profession	50
<b>REFERENCES</b>	52

<b>PART II. THE EARTH</b>	55
<b>5. EARTH AND ITS MOTIONS</b>	57
5.1. Earth's annual motion	57
5.2. Earth's spin, precession, and nutation	59
5.3. Earth's free nutation	63
5.4. Observed polar motion and spin velocity variations	66
<b>6. EARTH AND ITS GRAVITY FIELD</b>	70
6.1. Gravity field	70
6.2. Gravity anomaly	76
6.3. Gravity potential	82
6.4. Geoid and deflections of the vertical	87
<b>7. EARTH AND ITS SIZE AND SHAPE</b>	97
7.1. Actual shape of the earth	97
7.2. Geoid as a figure of the earth	104
7.3. Biaxial ellipsoid as a figure of the earth	110
7.4. Other mathematical figures of the earth	117
<b>8. EARTH AND ITS DEFORMATIONS IN TIME</b>	123
8.1. Tidal phenomena	124
8.2. Crustal loading deformations	130
8.3. Tectonic deformations	138
8.4. Man-made and other deformations	143
<b>9. EARTH AND ITS ATMOSPHERE</b>	151
9.1. Some physical properties of the atmosphere	151
9.2. Wave propagation through the atmosphere and water	155
9.3. Temporal variations of the atmosphere	161
9.4. Gravitational field of the atmosphere	164
<b>REFERENCES</b>	167
<b>PART III. METHODOLOGY</b>	173
<b>10. ELEMENTS OF GEODETIC METHODOLOGY</b>	175
10.1. General procedure	175
10.2. Formulation of the mathematical model	177
10.3. Observables and their properties	181
10.4. Vector of observables	188

<b>11. CLASSES OF MATHEMATICAL MODELS</b>	191
11.1. Classification of models	191
11.2. Models with a unique solution	196
11.3. Models with an underdetermined solution	198
11.4. Models with an overdetermined solution	200
<b>12. LEAST-SQUARES SOLUTION OF OVERRDETERMINED MODELS</b>	202
12.1. Formulation of the least-squares problem	202
12.2. Solution of the least-squares problem	204
12.3. Covariance matrices of the results	209
<b>13. ASSESSMENT OF RESULTS</b>	214
13.1. Hilbert space and statistics	214
13.2. Statistical testing	220
13.3. Assessment of observations of one observable	225
13.4. Simultaneous assessment of observations and mathematical models	231
13.5. Assessment of the determined parameters	239
<b>14. FORMULATION AND SOLVING OF PROBLEMS</b>	242
14.1. Optimal accuracy design	243
14.2. Analysis of trend	245
14.3. Adjustment of observations	258
14.4. Problems with a priori knowledge about the parameters	264
14.5. Problems with constraints and singularities	269
14.6. Step-by-step procedures in dynamic and static problems	276
<b>REFERENCES</b>	284
<b>PART IV. POSITIONING</b>	289
<b>15. POINT POSITIONING</b>	291
15.1. Fundamentals of geodetic astronomy	292
15.2. Astronomical positioning	304
15.3. Satellite positioning	309
15.4. Transformations of terrestrial positions	323
<b>16. RELATIVE POSITIONING</b>	335
16.1. Relative three-dimensional positioning	335
16.2. Relative horizontal positioning on reference ellipsoid	347
16.3. Relative horizontal positioning on conformal map	355
16.4. Relative vertical positioning	364

<b>17. THREE-DIMENSIONAL NETWORKS</b>	374
17.1. Three-dimensional networks using terrestrial observations	374
17.2. Photogrammetrical networks	381
17.3. Three-dimensional networks using extraterrestrial observations	385
17.4. Assessment and merger of three-dimensional networks	390
<b>18. HORIZONTAL NETWORKS</b>	396
18.1. Horizontal datum	396
18.2. Mathematical models and their solution	400
18.3. Assessment, expansion, and merger of horizontal networks	407
18.4. Marine positioning	416
<b>19. HEIGHT NETWORKS</b>	423
19.1. Vertical datum	423
19.2. Mathematical models for levelling	428
19.3. Assessment and design of height networks	438
19.4. Other heighting concepts	441
<b>REFERENCES</b>	447
<b>PART V. EARTH'S GRAVITY FIELD</b>	457
<b>20. GLOBAL TREATMENT OF THE GRAVITY FIELD</b>	459
20.1. Fundamental equations for gravity potential	459
20.2. Eigenfunction development of gravitational potential	467
20.3. Model gravity field	477
20.4. Disturbing potential	483
<b>21. LOCAL TREATMENT OF THE GRAVITY FIELD</b>	489
21.1. Conversion of disturbing potential into other field parameters	489
21.2. Vertical gradient of gravity	497
21.3. Curvature of the plumb line	503
21.4. Topographical and isostatic effects	508
<b>22. DETERMINATION OF THE GRAVITY FIELD FROM GRAVITY OBSERVATIONS</b>	516
22.1. Stokes's concept	516
22.2. Molodenskij's concept	526
22.3. Gravimetry	534
22.4. Evaluation of the surface integrals	539

23 DETERMINATION OF THE GRAVITY FIELD FROM OBSERVATIONS TO SATELLITES	546
23.1. Satellites and the gravitational field	546
23.2. Prediction of orbits	549
23.3. Analysis of orbital perturbations	554
23.4. Evaluation of gravity field parameters	559
24. DETERMINATION OF THE GRAVITY FIELD FROM DEFLECTIONS AND FROM HETEROGENEOUS DATA	564
24.1. Geometrical solution for the geoid	564
24.2. Transformation of gravity field parameters	569
24.3. Densification and refinement of deflections of the vertical	572
24.4. Solutions for the geoid from heterogeneous data	576
REFERENCES	581
 PART VI. TEMPORAL VARIATIONS	585
25. CORRECTIONS FOR TEMPORAL VARIATIONS	587
25.1. Elastic response to tidal stress	587
25.2. Tidal corrections	592
25.3. Corrections due to sea tide effects	600
25.4. Corrections due to polar motion deformations, and other causes	607
26. DETECTION OF VERTICAL MOVEMENTS	611
26.1. Sources of information on vertical movements	611
26.2. Interdependence of temporal variations of gravity and heights	615
26.3. Vertical displacement profiles	619
26.4. Areal modelling of vertical movements	625
27. DETECTION OF HORIZONTAL MOVEMENTS	633
27.1. Sources of information on horizontal movements	633
27.2. Comparison of horizontal positions	638
27.3. Direct evaluation of horizontal displacements	643
27.4. Strain, shear, and other models	649
REFERENCES	656
 AUTHOR INDEX	661
SUBJECT INDEX	669

PART I

# INTRODUCTION

## CHAPTER 1

### HISTORY OF GEODESY

Ever since man evolved into a thinking creature, he has been interested in learning about the earth. The various natural phenomena he observed around him, often with awe or fear, were frequently responsible for his behaviour and gave rise to various superstitions, rites, and cults. These, in turn, encouraged a better comprehension of events which resulted in many early cultures and civilizations acquiring a surprisingly deep understanding of some of the natural phenomena, left to us in such obvious forms as monuments (like Stonehenge in Wiltshire, southern England and the Egyptian pyramids), temples and towns (built by Central American Indians), calendars, etc. Such natural phenomena are often intimately related to the size, shape, gravity field of the earth, and their time changes, and to understand them requires a certain knowledge of geodesy.

For many centuries, the only way to learn about the geometry of the earth was through the observations of the sun, moon, planets, and stars, i.e., through astronomy. Hence the first achievements of geodesy went hand in hand with the development of astronomy. Together with astronomy, geodesy is among the oldest sciences; it is doubtless the oldest geoscience.

Little documentation of the geodetic accomplishments of the oldest civilizations—Sumerian, Egyptian, Chinese, Indian—has survived. There are many indications [TOMPKINS, 1971], however, that they must have had some very accurate observations, if not an understanding, of at least the basic motions of the earth. Our outline of geodetic history begins with the first positively documented concepts of the Greek era. Inevitably, the story we present is very subjective, with the historical flavour being emphasized rather than the historical accuracy. For the facts and dates not referenced in the text, the source is ASIMOV [1972]. Throughout this chapter we use modern terminology which, from the historical point of view, may at times be misleading. To do otherwise would have required more space than this presentation warrants.

This chapter is divided into four chronological sections. The first section covers the period from Thales till the end of the Roman Empire. The second section treats the Middle Ages, the Renaissance, and the beginning of the era of rationalism till about the mid-eighteenth century marked by the acceptance of Newton's theory of gravitation. The third section deals with the next 200 years, ending with the Second World War, and marked by the acceptance of Einstein's theory of gravitation. The last section describes the most recent developments of approximately the past 40

years. It was our deliberate decision to avoid references to living scientists, with a few notable exceptions.

### 1.1. Historical beginnings of geodesy

During the Greek era, geodesy was considered to be one of the most challenging disciplines and, consequently, some of the best intellects of that period devoted their energies to it. The first documented ideas about geodesy date back to Thales of Miletus (c. 625–c. 547 B.C.), commonly recognized as the founder of trigonometry. His concept of the earth was that of a disk-like body floating on an infinite ocean; our own interpretation of this idea is shown in FIG. 1.

Anaximander of Miletus (c. 611–c. 545 B.C.), Thales's contemporary, had a slightly different idea; he taught that the earth was cylindrical—see FIG. 2—with the axis oriented in the east–west direction [ASIMOV, 1972]. He was the first to discourse on a celestial sphere. This idea has permeated centuries of astronomical thinking and still remains a useful idealization in position astronomy (see §15.1). Anaximenes, Anaximander's pupil, modified Thales's vision somewhat by maintaining that the earth floated on a finite, circumferential ocean and was held in space by compressed air [BROWN, 1949]. This is interpreted in FIG. 3.

The school of Pythagoras (c. 580–c. 500 B.C.) was the first to believe in a spherical earth—a view that prevailed for well over two millenia. The work of this school was later compiled by Philolaus (flourished mid-fifth century B.C.) who was also the first to propose a non-geocentric universe centred on Hestia (the central fire). As the sun (and all the other bodies) moved in circular orbits round this fire, it cannot be called a heliocentric system [DIJKSTERHUIS, 1950]. Around the end of the sixth century B.C., Hecataeus of Miletus compiled one of the first known maps of the world, represented here in FIG. 4 (after BUNBURY [1883]). It rather vividly illustrates the limited knowledge and the prejudices the ancient Greeks had about the world. Yet,

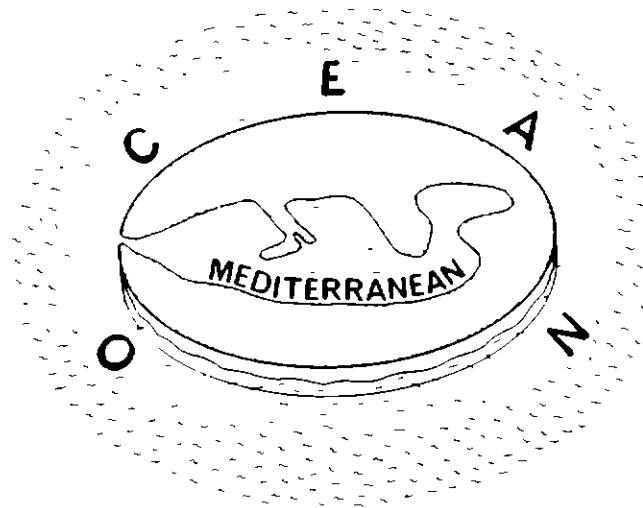


FIG. 1.1. Authors' interpretation of Thales's concept of the earth.

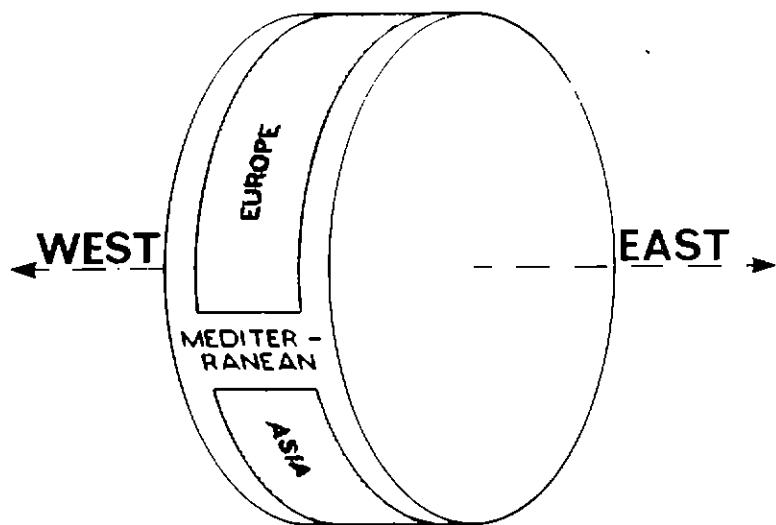


FIG. 1.2. Authors' interpretation of Asimov's description of the figure of the earth according to Anaximander.

at about this time, a Phoenician by the name of Hanno (born c. 530 B.C. in Carthage) may have circumnavigated Africa [WELLS, 1961]. As with the reports and findings of so many explorers throughout the ages, his were disbelieved and forgotten for another 2000 years.

Astronomy, often based not on observations but on philosophical views of the world, continued to develop. Anaxagoras (c. 500–428 B.C.) was the first to recognize the spherical form of the moon and explain the diurnal motions of the sun and the moon. The first star map was prepared by Eudoxus (c. 408–c. 355 B.C.) who also knew the length of the solar year almost exactly: 365.25 days, a figure probably learned from the Egyptians. Heracleides (c. 388–c. 315 B.C.) proposed that at least the earth, Mercury, and Venus moved round the sun, thus modifying Philolaus's century old notion. He also taught that the earth spins round its own axis.

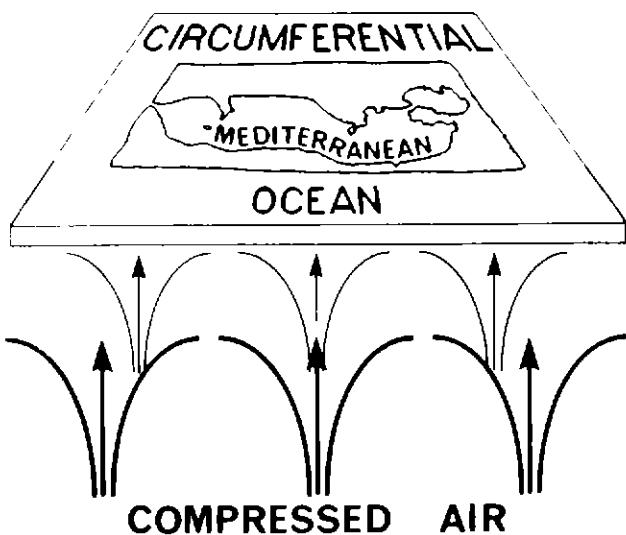


FIG. 1.3. Authors' modification of Brown's interpretation of Anaximenes's earth.

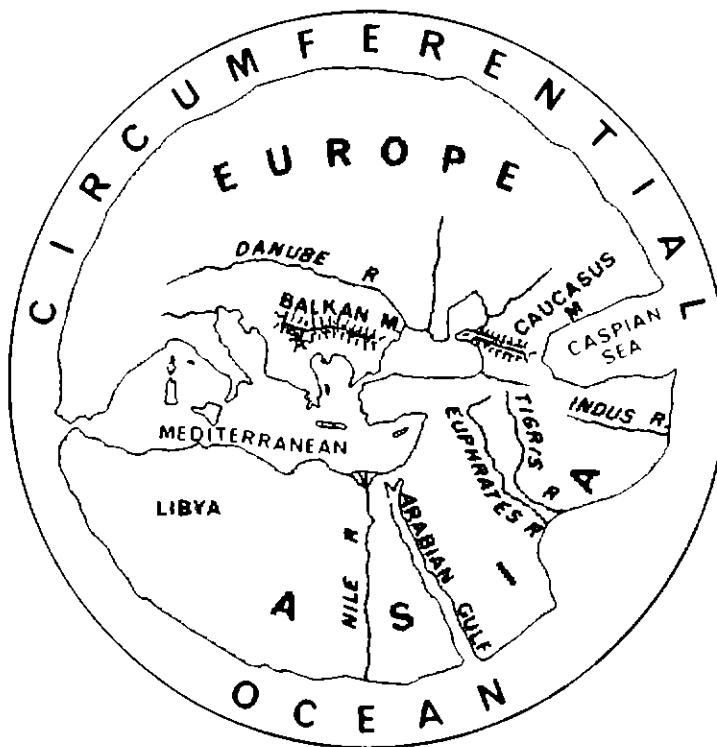


FIG. 1.4. Hecataeus's map of the world.

The first hint at the possibility of gravity is due to Aristotle (384–322 B.C.) who, in addition, formulated the first plausible argument for the sphericity of the earth, which survives till the modern day. Aristotle's interest in gravity was taken up by Strato (born c. 340 B.C.) after whom further breakthroughs had to wait till the Renaissance. Pytheas (born c. 300 B.C.) suspected the celestial bodies were responsible for the sea tide (see §8.1) but had insufficient knowledge to link this to gravitational attraction.

With the idea of the sphericity of the earth becoming acceptable, it was only a matter of time before spherical (angular) coordinates were introduced. This was finally done by Dicaearchus (died c. 285 B.C.) around the end of the third century B.C. He also compiled an updated map of the world containing information about south Asia gained during Alexander the Great's military expeditions. Shortly afterward, Pytheas determined the first relatively accurate latitude (for Marseilles).

Further progress in astronomy is associated with Aristarchus (c. 310–c. 250 B.C.) who attempted to determine the dimensions and distances of the moon and the sun. About half a century later, Eratosthenes (276–194 B.C.) introduced the notion of the obliquity of the earth's spin axis. Hipparchus (c. 190–c. 120 B.C.) gave us the first accurate star maps drawn in an angular system of coordinates, known now as the right ascension system (see §15.1). He subscribed to the idea of a precessing earth (see §5.2) but never accepted the heliocentric hypothesis of Heracleides, Aristarchus, and Seleucus, a Babylonian astronomer and contemporary of Hipparchus. It would be 1700 years before anyone again taught the heliocentric motion of the earth.

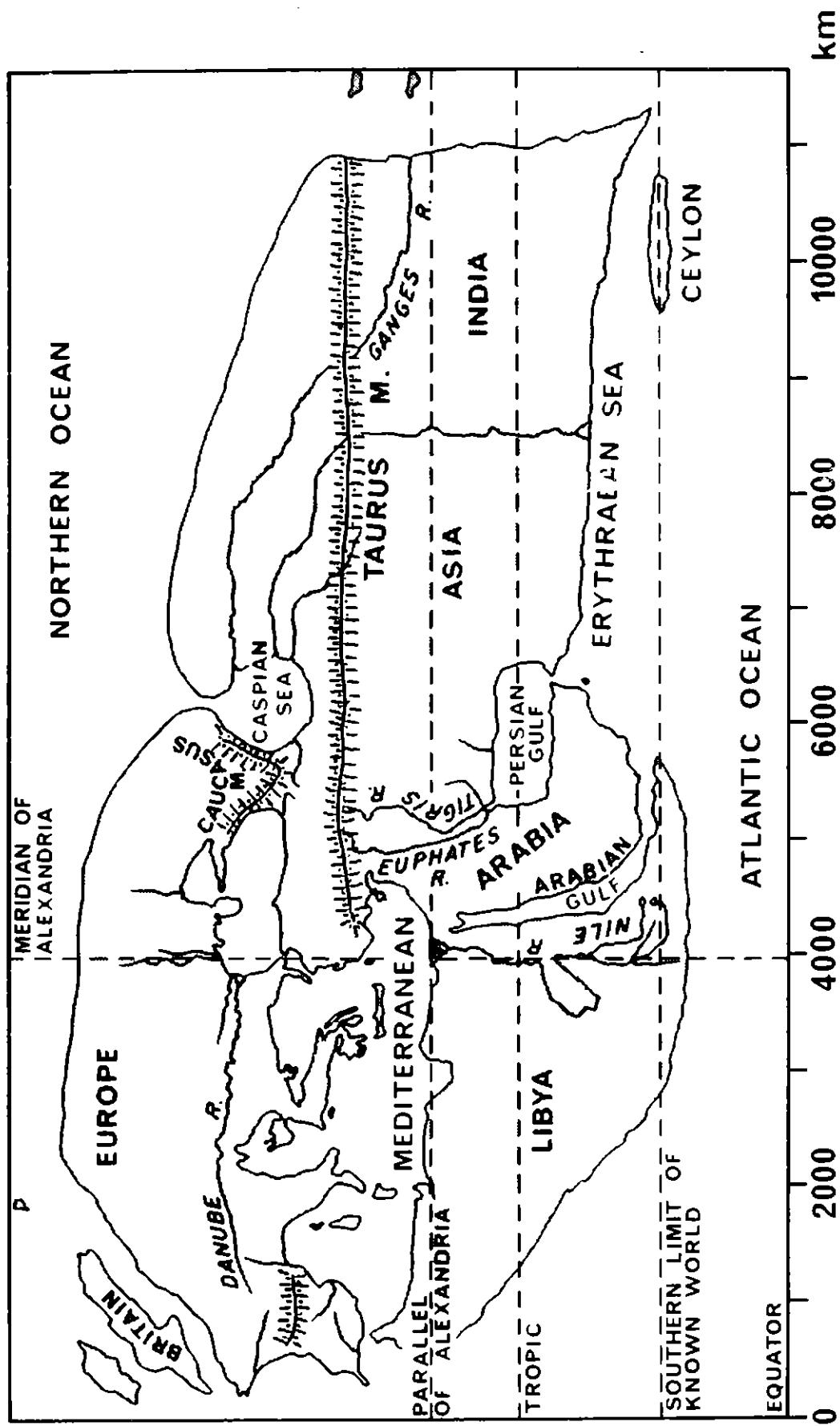


FIG. 1.5. The world according to Eratosthenes.

Let us return to Eratosthenes who, from the geodetic point of view, is the most interesting of all the aforementioned. Eratosthenes, holder of the prestigious position of librarian of the famous Alexandrian Museum (an institution approximating today's university), can be called the proper founder of geodesy. The result of his determination of the size of the (then thought of as spherical) earth, in his celebrated measurement of latitude difference between Alexandria and Aswan [GROUEFF, 1974], is discussed in §7.3 in the context of some of the more modern results. A later attempt by Poseidonius (c. 135–c. 50 B.C.), who considered the effect of air refraction (see §9.2), is now known to be considerably inferior to that of Eratosthenes. Along with some of his predecessors, Eratosthenes believed in the existence of one interconnected ocean which belief had to wait for confirmation for 17 centuries. His vision of the earth's surface is shown in FIG. 5 (after BUNBURY [1883]).

With Poseidonius there ended the era of original thinkers and experimenters. Thereafter, for some one and a half millenia, geodesy remained static, except for an occasional compilation or synthesis of the Greek achievements. The only notable exception during the Roman Empire was probably the implementation of the (Julian) calendar, commissioned of Sosigenes by Julius Caesar in the middle of the first century B.C. [DURANT, 1944]. This calendar, except for the small Gregorian reform in 1582 [PANNEKOEK, 1951], has survived till today.

As the Greek era drew to a close, some very important compilatory works were carried out by the Greek astronomer Claudius Ptolemy (c. 75–151). Ptolemy published a monumental compilation of astronomy and geodesy as developed at Alexandria, which is known under its Arabic name of *Almagest*. In an equally important work, the *Geography* published in 150, Ptolemy produced a new map of the world unsurpassed for some fourteen centuries. It is shown here in FIG. 6, according to THOMSON [1966]. It clearly represents no substantial improvement over the 300 year old map of Eratosthenes. In one respect it is worse: Ptolemy used the inferior figure of Poseidonius for his size of the earth at the expense of Eratosthenes's. An illustration of the intrinsic conservatism of the science of that period is the fact that Ptolemy never accepted the heliocentric hypothesis believed in by several astronomers before him. He also paid little heed to the traveller Strabo's (born c. 63 B.C.) suggestions that some continents could exist as yet unknown to man.

## 1.2. Scientific beginnings of geodesy

The ancients had been held back from expanding their knowledge of the material world by their philosophical and religious beliefs. In the centuries following the fall of the Roman Empire, i.e., during the Middle Ages, geodesy, along with so many other sciences, came more and more within the detrimental embrace of theology. The Greek teachings survived this dark period chiefly in Arabic versions that in the twelfth century found their way to Europe through Spain and were translated into Latin, then the language of European intellectuals. An example of the influence the Scriptures had on scientific thought in the European Middle Ages is shown in FIG. 7 (after BROWN [1949]), which is the navigator Cosmas's idea of the world of 548.

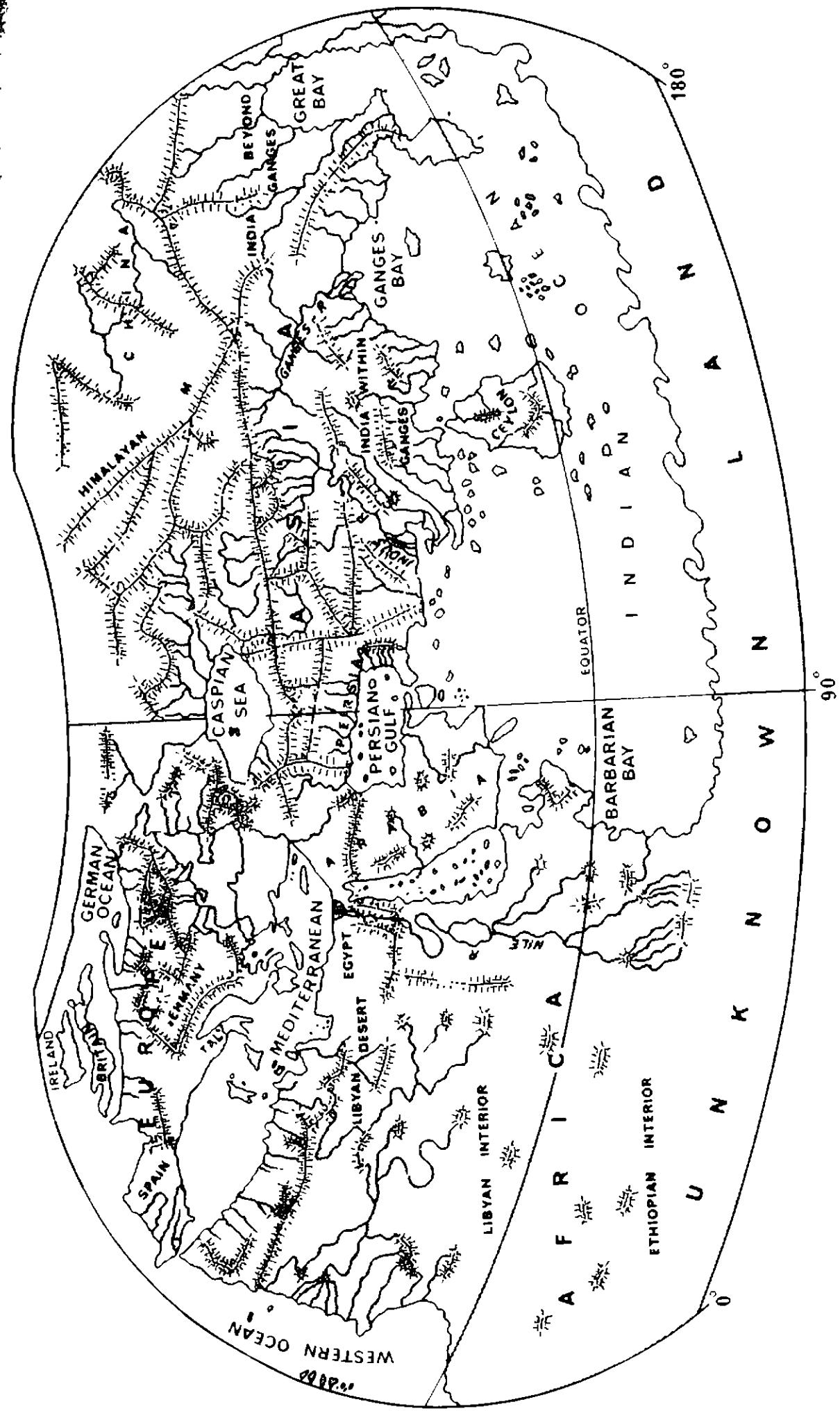


FIG. 1.6. The world according to Ptolemy.

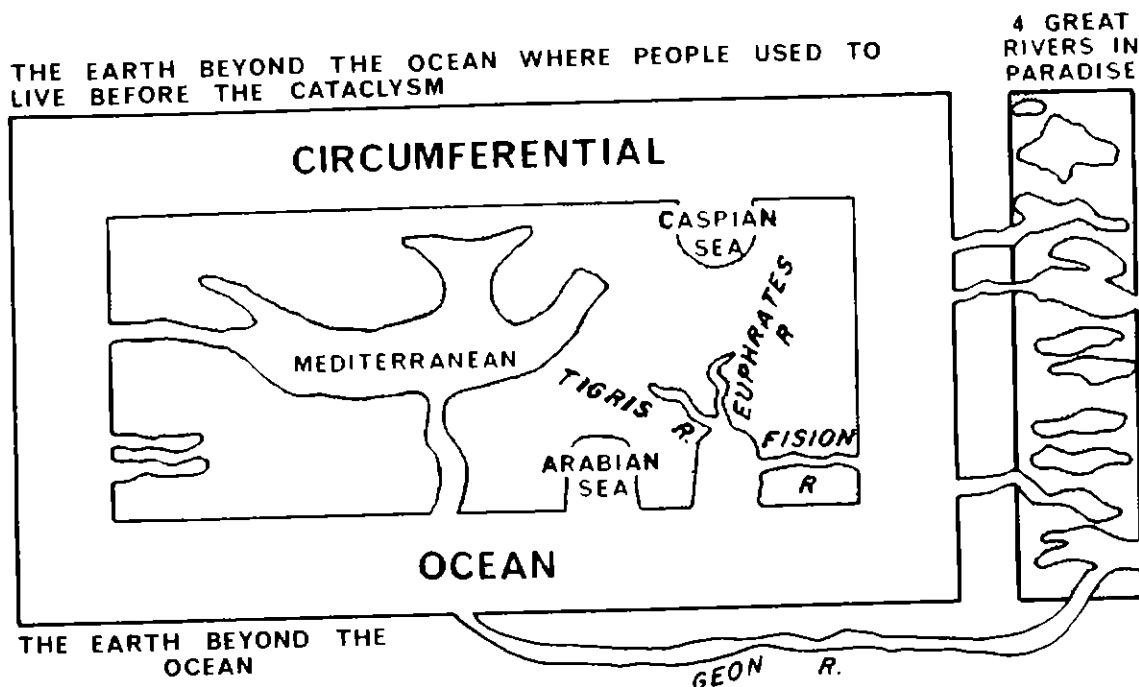


FIG. 1.7. Cosmas's vision of the world.

As will be seen from the following, the occasional glimpses of light during the Middle Ages were rare and far from overwhelming. The Persian Kharazmi (born c. 780), after whose Arabic name, Al-Khwarizmi, comes the word 'algorithm', re-determined the size of the earth. The result was about 1.6 times too large—no match for Eratosthenes's. Al-Khwarizmi, who also published a map of the world not very different from Ptolemy's, earns a permanent place in history by introducing Hindu numerals, 1, 2, ..., 9, into Arabic mathematics. The Arabic astronomer Albategnius (c. 858–929) knew the length of the year more accurately than Sosigenes did nine and a half centuries earlier. So did the Englishman Roger Bacon (c. 1210–92), who advocated reforming the Julian calendar to include one extra day every 128 years.

Things started coming to a head in the mid-fourteenth century, characterized by renewed curiosity and growing boldness. The age of great explorations was approaching, and the quest for uncorrupted truth grew. A new vision of the world, doubtless influenced by the exploits of Marco Polo (in the period 1271–95), was offered by Toscanelli (1397–1482) and is shown here in FIG. 8 (after HAPGOOD [1966]). It was reputedly this map and Bacon's estimate of the short distance from Europe to the east coast of Asia that tempted Columbus to sail west to find the new, only 5000 km long, way to India [DURANT, 1944].

The major explorations got under way at the end of the fifteenth century with Columbus crossing the Atlantic in 1492, Vasco da Gama circumnavigating Africa in 1497, and Magellan's expedition circumnavigating the world between 1519 and 1522. The expanding geographical knowledge prompted the growth of a new profession: map-making, or cartography. Cartography is the art of displaying the final product of geodesy, so mention must be made of a few of the more famous map-makers in

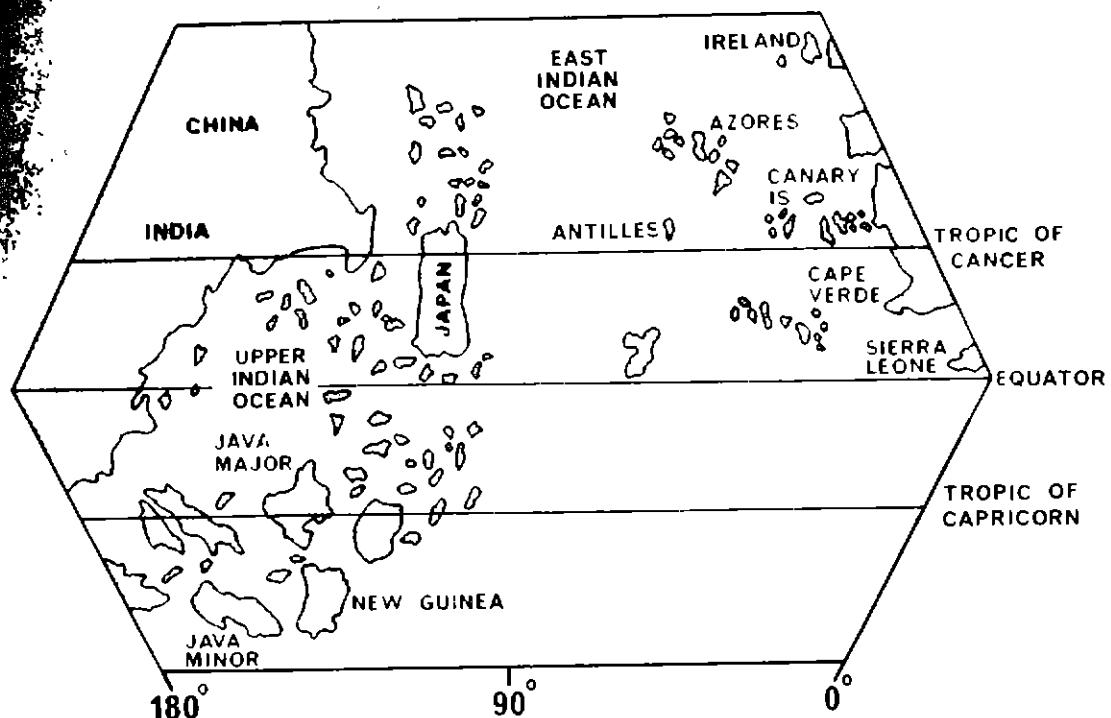


FIG. 1.8. Toscanelli's idea of the western hemisphere.

history. Among the best known is the Italian Amerigo Vespucci (1451–1512) who gave us the first maps of the North American Pacific coast and provided a name for the continent. Another well-known map-maker, often considered to be the father of modern cartography, is the Flemish Mercator (1512–94). He very successfully responded to the demands of navigators for maps with the least distortions (see §16.3). FIG. 9 (after FITE AND FREEMAN [1926]) shows one of his world maps which reflects the tremendous improvement, during the Renaissance, in mankind's knowledge of the earth's surface. Although Eratosthenes's figure for the size of the earth was finally accepted after it had been confirmed by Magellan's expedition, old customs die hard, and maps like the one shown in FIG. 10 (after NORDENSKJÖLD [1889]) were still being printed in the mid-sixteenth century.

Indications of an impending revival of geodesy can be found in the mid-fifteenth century when there arose a series of thinkers who paved the way for Copernicus and Kepler. Among the better known were the German cardinal Nicolaus of Cusa (1401–64), who wrote about the diurnal motion for the earth and introduced the idea of an infinite universe, and the Italian artist Leonardo da Vinci (1452–1519) who suggested the probability of isostasy (see §8.2) [DURANT, 1944]. Finally, about 1530 the Polish clergyman Copernicus (1473–1543) published his heliocentric theory which, for the first time, included all the planets.

The battle of reason against theology though was not over. In 1600 the Italian astronomer Bruno (1548–1600) died at the stake for, among other heresies, maintaining basically the same views that Nicolaus of Cusa and Copernicus had held before him. The story of Galileo's forced recant [WELLS, 1961] (an apology finally being issued by Pope John Paul II in November, 1979) of heliocentricity is well

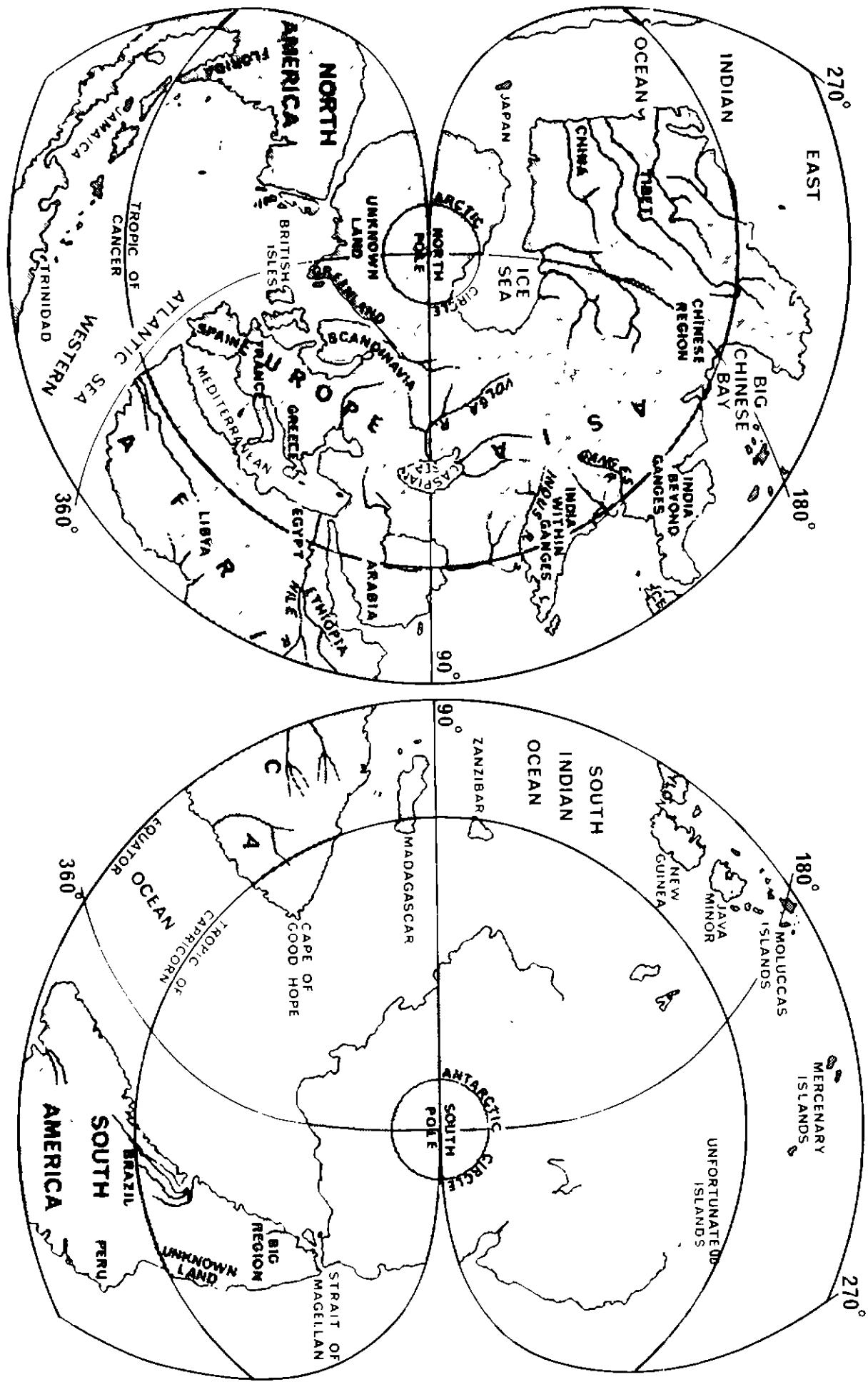


FIG. 1.9. Mercator's map of the world.

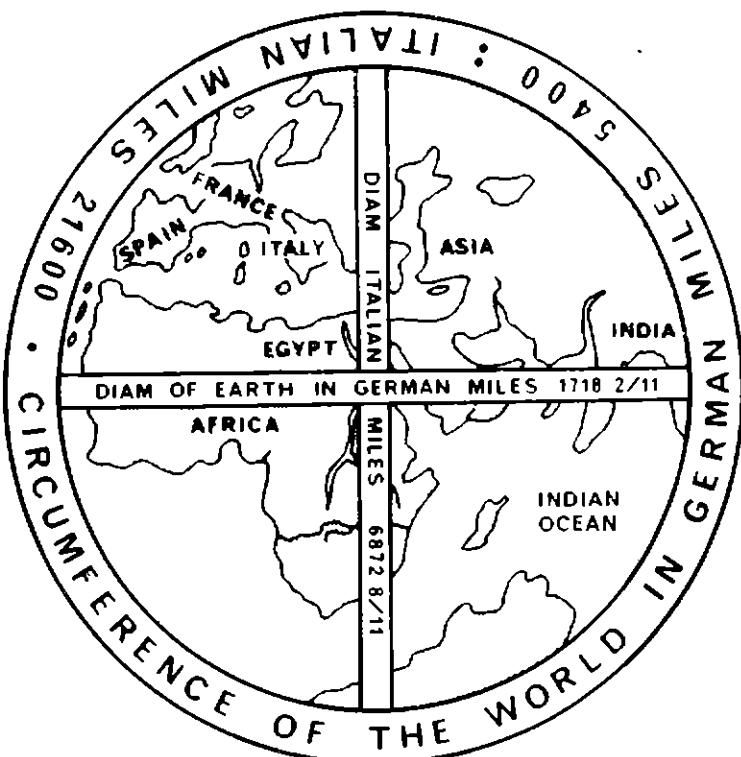


FIG. 1.10. Apianus's map of the world.

known. The observational evidence, collected chiefly by the Danish astronomer Tycho Brahe (1546–1601); improvements of the experimental methods, due mainly to the Italian Galileo (1564–1642); progress in theory, associated with the German Kepler (1571–1630); and superior instrumentation (such as the telescope) should have combined to render the theologically based views untenable. But in Catholic countries the Inquisition banned the books of Copernicus, Kepler, Galileo and others advocating heliocentricity until 1822 when they were finally removed from the *Index* [DREYER, 1905].

In the meantime for geodesy, this fermentation of ideas brings the beginning of real scientific enquiry into gravity in terms of the Dutchman Stevin's (1548–1620) experiment showing the equivalence of gravity attraction for disparate bodies, and Galileo's formulation of the first mechanical laws. Nevertheless, Newton's idea of gravity as a force was still far away. In 1615, the Dutchman Snell (1591–1626) carried out the first accurate triangulation [BÖHM, 1972] and made the first rigorous study of refraction. The French clergyman Picard, in 1670, made the first modern measurement of the size of the earth. His result of 6275 km [GROUEFF, 1974] for the radius of the earth is the first improvement on Eratosthenes in 19 centuries. The technique Picard used is outlined in §7.3.

The scene was set for the most important discovery of this era: the (Newton's) law of universal gravitational attraction in 1687 (see §6.1), for which the works of the Italian Borelli (1608–79) and the Englishman Horrox (1619–41) can be seen as precursors. The mathematical tools needed had been prepared by Descartes (1596–1650), Leibnitz (1646–1716), and Newton (1642–1727) himself who, among other

things, was a professor of mathematics at the University of Cambridge. Progress in the understanding of gravity brought about two somewhat related discoveries. Towards the end of the seventeenth century, the Dutchman Huygens invented the first accurate time keeping mechanism based on the use of a pendulum; the Englishman Bradley (1693–1762) discovered nutation (see §5.2).

Newton's theory of gravitation was not accepted overnight. Its most renowned opponent was Newton's French counterpart, the royal astronomer of Italian origin, Cassini (1625–1712). While Newton's new theory predicted that the earth should be oblate—because of the centrifugal force caused by the spin (cf. §6.1)—Cassini maintained that it should be prolate. This he did in spite of the observational discovery by the Frenchman Richer in 1671 that gravity was weaker at the equator, as required by Newton's theory.

As the theory of gravitation gained acceptance, a resolution of the deadlock between Newton and Cassini was required. In the years 1735–43, the French Academy of Science organized two survey expeditions to measure two meridian arcs—and the corresponding latitude differences—one at the equator, the other closer to the pole. The equatorial expedition, under the leadership of Bouguer, went to Peru (now Ecuador). The other, led by Maupertuis (1698–1759) and including the young Clairaut (1713–65), went to Lapland. The results of the two expeditions confirmed the validity of Newton's theory. In addition, it was Clairaut who, as a by-product of his theory of rotating fluid bodies, later derived the simple relationship between the gravity change along a meridian and the flattening of the earth (cf. §7.4).

### 1.3. Geodesy in the service of mapping

The pioneering work done by Snell, Picard, and the two French expeditions showed that terrestrial geodetic measurements (angles and distances) are viable tools for the task of relative positioning. Networks of points whose horizontal positions were determined from the measurements of angles and occasional distances (see §7.1), known as triangulation networks, started to spring up in all parts of Europe in support of mapping programmes of various kinds. Accurate mapping for military as well as civilian purposes became feasible because it was suddenly possible to cover the land with triangulation points the positions of which were obtainable with relative ease. The instruments needed for triangulation, i.e., theodolites and base line measuring devices like wires and tapes, became more precise, easier to operate, and more portable. The techniques of triangulation, astronomical determination of positions and azimuths, as well as levelling, have been perfected (cf. Part IV). Between 1750 and 1950, the determination of positions from terrestrial and astronomical observations were the daily bread of geodesists. So much so that even today many people view geodesy as being merely a synonym for this task.

At times, these geodetic tasks presented an intellectual challenge to the best brains of the era, arousing an interest equal to that which geodesy stirred at the dawn of our civilization. Thus we find, for instance, J. K. F. Gauss (1777–1855), acknowl-

edged as the greatest mathematician of the early nineteenth century, inventing the heliotrope, a device that uses reflected solar rays for signalization of geodetic points, and measuring a geodetic network in the kingdom of Hannover. In America, with its smaller population density and larger distances, unique techniques had to be used by surveyors (like George Washington) to meet the more challenging problem of positioning. The first satisfactory map of British and French North America became available in 1755 [BOORSTIN, 1958].

Hand in hand with developments in geodetic positioning went discoveries in other aspects of geodesy. In 1798 the Englishman Cavendish, using Michell's torsion balance, succeeded in 'weighing the earth'. The French mathematician Laplace (1749–1827) laid the foundations for modern celestial mechanics and the theory of tides; he also devoted a considerable effort to the development of probability theory. The German astronomer Bessel (1784–1846) determined the first accurate figure of the flattening of the earth (see §7.3) from existing knowledge of geodetic positions. Gauss defined the geoid (see §6.3) and invented the least-squares method (see Chapter 12), though concurrently with Legendre. His work on the theoretical foundations of geodesy has caused some geodesists to claim him as the father of geodesy. He did usher geodesy into its mature age, but he was equally eminent in other branches of science.

The end of the eighteenth and the whole of the nineteenth centuries were enormously fruitful in the realm of mathematics. Most of the tools of applied mathematics used in geodesy today were invented then. Thus mention should be made of a few great mathematicians who contributed the most toward building up the geodetic 'arsenal'. These are: the Swiss Euler (1707–83), with his work on mechanics of physical bodies; the French-Italian Lagrange (1736–1813), the creator of analytical mechanics who, among other contributions, helped to introduce the metric system in France in 1795. Another Frenchman, Fourier (1768–1830), is remembered for his work on potential, Gauss and the German Riemann (1826–66) for their work on differential geometry, and the Irishman Hamilton (1805–65) who put the finishing touches to analytical mechanics.

Naturally, in this period of rationalization, other fields akin to geodesy underwent equally fast development. To name a few: geophysics began with the Scottish geologist Hutton's (1726–97) theory of evolution of the earth's surface, the German polyhistor Humboldt's (1769–1859) studies of various physical aspects of the earth, and the German geophysicist Wegener's (1880–1930) theory of continental drift (see §8.3). The elevation determined by Humboldt of Chimborazo in South America [BOTTING, 1973] remained the highest known until the measurements in the Himalayas started by Everest, the Surveyor General of India. Oceanography progressed from the first soundings carried out by the English explorer Cook (1728–79), to the American oceanographer Maury's (1806–73) mapping of the sea bottom and the currents, to the Swiss explorer Piccard's (1884–1962) observations from submersibles. Propagation of electromagnetic waves was theoretically described by the Scottish physicist Maxwell (1831–79), and its velocity first measured in a laboratory by the Frenchman Fizeau (1819–96). The application of electromagnetic waves to long distance measurements was carried out by the German-American physicist

Michelson (1852–1931) who first determined a geodetic distance to a relative accuracy better than  $10^{-6}$ .

All these developments had a stimulating effect on geodesy, and discoveries in the realm of geodesy proper followed. The French physicist Coriolis (1792–1843) explained the total acceleration of bodies moving on the earth's surface. The mid-nineteenth century saw the first measurements of the deflections of the vertical (see §6.4) and the first attempts by two English physicists, Airy and Pratt, to quantify isostasy (§8.2). At about the same time, the French physicist Foucault demonstrated that the earth is spinning and invented the gyroscope, later to be adapted into a gyrocompass (see §16.1) by the American Sperry (1860–1930). The year 1880 saw the first serious attempt to synthesize and formalize the mathematical and physical foundations of geodesy by the German geodesist Helmert in his book *Mathematical and Physical Theory of Geodesy*. In 1883, the English physicist Stokes published the solution of the geodetic boundary value problem (see §22.1) in closed form. The Scot Kelvin (1824–1907), the Englishman Darwin (1845–1912, son of Charles Darwin), and the Frenchman Poincaré (1854–1912) developed the theory of the earth tides (see §8.1), and the Canadian astronomer Newcomb (1835–1909) studied the wobble of the earth's spin axis (see §5.4).

The beginning of the twentieth century saw a major change in the thinking of physicists affected by Minkowski's space-time and, of course, by Einstein's special and general theory of relativity [CLARK, 1971], a further generalization of Newton's theory of gravitation. The idea that "... gravity is geometry—the geometry of space and time..." [DAVIES, 1979] soon permeated physics and, though not directly applicable to most geodetic problems, had an effect on geodesy in due course. It has certainly affected at least the philosophical outlook of the authors of this book.

In the first half of the twentieth century, the Hungarian physicist Eötvös studied gravity gradients, and the Dutch geophysicist Vening Meinesz significantly improved the theory of isostasy. The English geophysicist Jeffreys introduced the concept of the telluroid (see §7.4) that started a new trend in geodesy culminating in the Russian physicist Molodenskij's more rigorous solution to the geodetic boundary value problem (see §22.2). Finally, the work of the Italian mathematicians, Pizzetti and Somigliana, on the theory of the normal gravity field (see §20.3) must be mentioned.

#### 1.4. Geodesy of the modern era

The mid-twentieth century saw the dawning of the technological revolution. Prompted by weapons and defence requirements during the Second World War, the invention of a 'radio detection and ranging' system, commonly known as radar, has had a deep effect on the philosophy behind geodetic instruments. At about the same time, the first practical electronic computers appeared, opening up horizons for numerical mathematics unimaginable in the past. The introduction of computers not only sped up geodetic computations but revolutionized the thinking of geodesists:

solutions to tasks, previously out of the question because of the sheer volume of the calculations involved, became not only feasible but even easy.

For centuries, horizontal angles, measurable to a much higher accuracy with intrinsic ease, had been preferred to distances. Shortly after the war, sufficiently accurate electromagnetic distance measuring devices became commercially available for geodetic uses. These instruments, first using polarized light then radiowaves and finally lasers, eventually changed the pattern of geodetic positioning.

The forerunners to the turbulent development of extraterrestrial methods were the first experiments in radio-astronomy that culminated in the discovery of pulsars and quasars. These new distant radio-objects emit signals with high frequential stability and are now being used in the fast developing techniques of radio-interferometry (see §16.1).

The launching of the first artificial satellites was another giant leap for geodesy. For the first time, geodesists could use extraterrestrial objects, passive or active, for accurate positioning of points the intervisibility between which was no longer a constraint. The low altitude of the satellites offered the opportunity of studying the geometry of the earth's gravity field by means of direct observations of the satellite response (motion) to the field (cf. Chapter 23). Satellites also brought about a new project for geodesy: the mapping of the gravity field above the earth to predict satellite orbits. Once again, the major customers for this kind of information were the military who needed to know the gravity field geometry for computing missile trajectories.

Another spin-off of the space programme is the inertial navigation and positioning systems (see §16.1). These technologically complex systems were made possible by vast improvements in the accuracy of acceleration sensing and direction seeking devices. The spectacular development of microelectronics was probably the single most important contributor here.

The increased ease and accuracy with which geodesists could determine positions, as well as the gravity field parameters, led to new applications, but also to new problems. Suddenly, effects that had always been considered negligible started showing up, and the 'noise' these effects caused had to be accounted for. "One man's noise being the other man's signal...", other disciplines became interested in geodetic techniques, as well as results, to study the phenomena relating to their own fields. Prime examples of such (symbiotic) relations are those of geodesy with geophysics, space science, astronomy, and oceanography (see §2.2).

The relation with geophysics has been particularly fruitful because of another fact: in the late 1960s the hypothesis of plate tectonics finally gained almost universal acceptance. In some parts of the world (cf. §8.3), the rate of relative tectonic movement is so fast that it is directly measurable by geodetic means. Geodesy, therefore, became the major supplier of geometrical information on these movements. This successful deployment of geodesy in tectonic investigations has led to further applications of geodetic techniques in other branches of geodynamics.

The last important development of geodesy that must be mentioned here concerns the sea. Expansion into the marine environment, characterized by exploration and

exploitation of resources on the sea bottom, presented a new task to geodesists: positioning of moving as well as stationary objects on the seas. Part of the role geodesy plays in the marine environment is helping to satisfy the steadily growing demand for accurate navigation.

It is this rejuvenated geodesy, acquiring new dimensions, facing new tasks, and being provided with new techniques and tools, that we try to present in this book.

## CHAPTER 2

### GEODESY AND OTHER DISCIPLINES

It is beneficial to any student of geodesy to know the relationship of geodesy to other disciplines. It is these relations that determine the degree of usefulness and acceptability of any field of endeavour and, in the last analysis, dictate its scope. This chapter is our attempt to trace the connections geodesy has with other sciences and fields as we see them. These links vary from country to country as well as with time. Thus our classification is, of necessity, subjective, and doubtless there are geodesists who will disagree with our views. We make no attempt to classify the fields outside geodesy as to their hierarchy and interrelations. We do try, however, to show the flow of information vis-à-vis geodesy; this flow is used as our means of classification.

The first section discusses the disciplines in which geodesy is being applied. The second section briefly describes the symbiotic relation of geodesy with some other sciences. The third section focuses on those disciplines which provide the scientific basis for geodesy. Among these, mathematics plays the dominant role; so dominant, in fact, that a whole chapter (3) is devoted to those parts of mathematics commonly used in geodesy. Contrary to our inclinations, we have had to start with the applications and finish with the basics to ensure a logical continuation of the presented material from chapter to chapter.

#### 2.1. Applications of geodesy

Before examining the applications of geodesy, let us clarify the relation between geodesy and surveying: in most languages, no real distinction is made between the two. The distinction inherent in the English language probably causes more problems than it solves. Be that as it may, in our view *surveying* is the practice of positioning, and geodesy is the theoretical foundation of surveying.

For centuries, the role of geodesy was to serve mainly mapping (see §1.3)—an end many people still regard as the major purpose of geodesy. This reduction of geodesy to *control surveying*, whose sole function is to provide position control for mapping, simply is not warranted. Although a significant part of the information provided by geodesy falls within the realm of positioning, an equally substantial contribution is made elsewhere (see §4.1).

Let us now turn to the disciplines where geodetic information, positions or other, is needed. Our survey has been directed by the ideas contained in the following publications: KRAKIWSKY AND VANIČEK [1974], VANIČEK [1976], (U.S.) COMMITTEE ON GEODESY [1978], MUELLER [1978], HIEBER AND GUYENNE [1978], and RINNER [1979].

(a) *Mapping*: It is well understood that there is a need for an areal network of appropriately distributed points (geodetic control) of known horizontal and vertical positions for the production of maps ranging from small scale maps of entire countries to large scale maps used by municipalities. The establishment of this control is clearly an important geodetic task and will be discussed in Part IV.

(b) *Urban management*: In the urban environment, the locations of man's creations, such as underground utilities, must be defined and documented for future reference. The need for control points is clearly indicated in the literature, e.g., see BLACHUT ET AL. [1979].

(c) *Engineering projects*: During the building of large structures, such as dams, bridges, and factories, it is necessary to lay out the various components of these structures in predetermined locations. For this purpose, coordinates of one kind or another are used, so the availability of control points is naturally desirable. As well, it is often necessary to know the movements of the ground and water levels prior to, during, and after the construction. In the case of dams, water tunnels, irrigation projects and the like, the exact shape of the equipotential surfaces of the gravity field should be known. The determination of the movements (see Chapters 26 and 27) and the shape of the equipotential surfaces (see Part V) are also geodetic undertakings.

(d) *Boundary demarcation*: The rigorous definition of international and intranational (provincial or state) boundaries is of paramount importance. Emphasis has also recently been placed on the speedy and accurate description of oil and gas leases, even in such remote and inhospitable parts of the world as the arctic, the North Sea, and various continental shelves. The positioning and staking out of these boundaries is most economically done by relating them to a framework of points with known horizontal coordinates—the geodetic network (see Chapter 18).

(e) *Ecology*: It has been realized in the past few decades that it is necessary to study the effects of human actions on the environment. One such effect is the movement of the ground caused by the removal of underground resources (including water, oil, and minerals) or subsurface disposal of wastes (cf. §8.4). The detection and monitoring of these movements is a geodetic problem (see Chapters 26 and 27).

(f) *Environmental management*: It has been recognized that the establishment of environmental data banks, to serve as integrated information systems for transportation, land use, community and social services, land titles extracts, assessment of tax data, and population statistics, should be based on land parcels whose locations are uniquely defined in terms of coordinates [HAMILTON, 1969]. Again, it is advisable that these coordinates be referred to a geodetic network.

(g) *Geography*: All the positional information needed in geography is provided by geodesy. Even though the accuracy of positional and other geometrical information used by geographers is generally much lower than that needed in the fields described above, this information is of a global character that only geodesy can satisfy.

(h) *Planetology*: It can be argued that this is a part of either astronomy or geophysics. Be that as it may, planetology uses methods for studying the geometry, gravity fields, and deformations of planets that are identical with the extraterrestrial methods used in geodesy. Thus, practically all of geodesy is applicable to planetology. Because of this special affinity between planetology and geodesy, some geodesists regard the determination of the shape and size of planets and their gravity fields as part of geodesy.

(i) *Hydrography*: Some consider this field to be a part of oceanography, while others make it a special (marine) branch of surveying; either way, it has a somewhat special relation with geodesy. It may be understood as the practice of positioning at sea, combined with depth sounding, and, as such, applies many geodetic methods (see §18.4 and §19.4).

## 2.2. Symbiotic relation between geodesy and some other sciences

Clearly, there are many more uses for geodesy than simply mapping. Still other applications of geodesy are found in scientific fields that have a two-way relation with geodesy: while geodesy supplies one kind of information to them, they provide another kind of information for use in geodesy. Such fields are as follows:

(a) *Geophysics* has a history of probably the closest affiliation with geodesy. So much so that in some countries geodesy is regarded as a branch of geophysics. Because of this close relationship, it is sometimes difficult to distinguish where geophysics ends and geodesy begins: the boundaries are somewhat blurred. Consequently, it is to be expected that our account here will not be shared by all geodesists and geophysicists.

Geophysics, along with many other fields, requires the positions and other geometrical information geodesy can supply. In particular, it needs the geometrical information on the earth's temporal deformations. As explained in §1.4, geodetic techniques are being used increasingly in the detection of tectonic movements (see, e.g., SAVAGE AND BURFORD [1973], NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (NASA) [1979]). In other parts of contemporary geodynamics as well, geodetic data are used to obtain the geometry of the deformations [VANÍČEK, 1977].

Gravity is one of the most important sources of information used in both theoretical and *exploration geophysics* [TELFORD ET AL., 1976]; gravity data is necessary for studying the irregularities in the (underground) mass density distribution. Since geodesists are also vitally interested in gravity data to study the geometry of the gravity field (see §4.1), both sciences claim a jurisdiction over gravity data collection (gravimetry). A somewhat artificial division would assign global gravity work to geodesy, while regional and local gravity measurements would be a geophysical task. The temporal variations of the gravity field offer a valuable hint about the physical causes of vertical crustal movements. As such, these data are often exploited in the context of contemporary geodynamics (see §26.1).

In return, geophysics offers an insight into the physical response of the earth to a variety of forces (cf. Chapter 8), into the possible density distribution within the

earth (cf. §6.1), and into the effects of the internal structure of the earth on its motion (cf. §5.3). This information is needed when various mathematical models (relations) for geodetic purposes are being designed.

(b) *Space science*, compared with geophysics, is a very young field. Right from the beginning, its relation to geodesy has also been a very close one. The main reason is that the knowledge of the geometry of the earth's external gravity field is essential for predicting the orbits of space vehicles (cf. §23.2). In addition, locations of satellite tracking stations must be known precisely enough to be of use [NASA, 1978]; these are determined by geodetic means.

On the other hand, space science has developed some very powerful positioning systems that use the earth's artificial satellites, and these are now being used in geodesy to complement the existing terrestrial techniques (see Chapters 15 and 16). The analysis of the observed close satellite orbits now provides the best long wavelength data on the earth's gravity field, including the value of the flattening of the earth (cf. §7.3 and §23.4). Tracking of deep space probes gives the best estimates of the value of the mass of the earth.

(c) *Astronomy*, one of the oldest sciences in existence, and geodesy developed hand in hand for a long time (cf. Chapter 1). Although the interdependence of geodesy and astronomy has somewhat diminished in the recent past, positional visual astronomy still plays a certain role in geodesy (cf. §15.1 and §15.2). In addition, the future will probably see an increasing involvement of positional radio-astronomy (see §16.1). Another part of astronomy, celestial mechanics, is also needed in geodesy to study the satellite orbits (see §23.1). Geodesy shares with astronomy the interest in lunar laser ranging (see §16.1); the ranges are used in astronomy to compute the lunar orbit and libration [COOK ET AL., 1977], while geodesists use them for position determination. Of common interest too is the monitoring of the rotation of the earth (cf. §5.4).

(d) *Oceanography* is another science with which geodesy shares interests. Both geodesy and oceanography are involved in the location and movements of shore-lines. Geodesy provides the oceanographers with relative heights of the on-shore water level measuring devices (tide gauges) and their relative vertical movements [LENNON, 1974]. Also, the geodetically determined positions (see §18.4) of various marine objects, including ice and oceanographical vessels, are of value to oceanographers.

Oceanographical information of interest to geodesists includes the dynamics of the sea surface (cf. §8.4) and the deviations of the mean sea surface from an equipotential surface of the earth's gravity field (see §7.2). This information is needed for the establishment of a datum for heights.

(e) *Atmospheric science*, along with all the aforementioned sciences, uses the geodetic positions and gravity pertaining to meteorological stations and probes. It shares with geodesy an interest in satellite orbit analysis: while geodesy interprets the orbital perturbations in terms of gravitational effects, atmospheric science looks at the effect of the distribution of air density [JACCHIA AND SLOWEY, 1975]. Geodesy needs realistic models for atmospheric refractivity and the appropriate meteorological data to evaluate atmospheric refraction (see §9.2), which represents one of the most troublesome problems in many geodetic measurements (as will be seen in Part

V). Meteorological data are also needed in the analysis of sea level temporal variations (cf. §8.4 and §19.1) and, in special cases, that of the temporal variations of the earth's surface (cf. §8.4).

(f) *Geology* requires both horizontal and vertical positions for its maps. In return, it provides geodesists with a knowledge of geomorphology and the local stability of different geological formations. The information on stability is a must for any geodesist in charge of selecting suitable sites not only for geodetic monuments (see §7.1) but also for observatories of various kinds.

### 2.3. Theoretical basis of geodesy

The last group of disciplines to be mentioned are those providing the theoretical basis for geodesy. Representing a fairly standard foundation for many sciences, they are mathematics, computer science, and physics.

(a) *Mathematics* is, by far, the most important building block of geodesy. In fact, some sources regard geodesy as a branch of applied mathematics (see, e.g., ENCYCLOPAEDIA BRITANNICA [1970]). There is something to be said for this notion, since geodesy is essentially geometry applied to the earth. To do mathematics the justice it deserves in the context of geodesy, the whole of Chapter 3 is devoted to a description of the mathematical concepts needed in geodesy—at least the geodesy as presented in this book. It should be stated here that, while we have elected to include statistics in our treatment of mathematics in Chapter 3, numerical analysis is treated under the heading of computer science.

(b) *Computer science* teaches us how to use the computer systems, the most powerful computing and analytical aid available to us. Many of the problems faced by geodesy today require a computer solution. Geodesists, like most other scientists, should have an appropriate knowledge of at least one high level programming language and be adequately familiar with the interactive and graphical capabilities of a computer. Because of the large quantities of data involved in most geodetic problems, geodesists should have a proper training in data handling.

Last but not least, various numerical analysis concepts are needed in geodesy. Foremost are those pertaining to approximation theory, treated to a certain extent in §14.2. Numerical methods of linear algebra, discussed intermittently in Chapters 12 and 14 and used throughout the book (notably in §18.2), are a must. Also useful are numerical integration, differentiation, and quadrature of differential equations.

(c) *Physics* is almost as important to a geodesist as is mathematics. Since Newton, gravitation (see Chapter 6 and Part V) has played a very important role in geodesy; this importance only increased when it was realized that gravity is the geometry of the space in which most geodetic observations are taken (cf. §1.3). Today, the geometry of the earth's gravity field is considered a part of geodesy (see §4.1) as opposed to physics.

Of similar, fundamental importance in geodesy is the theory of propagation of electromagnetic waves. Almost all geodetic instruments use the principles of this propagation one way or another, and an understanding of the physical laws governing propagation (see §9.2) is thus essential to our comprehension of the nature

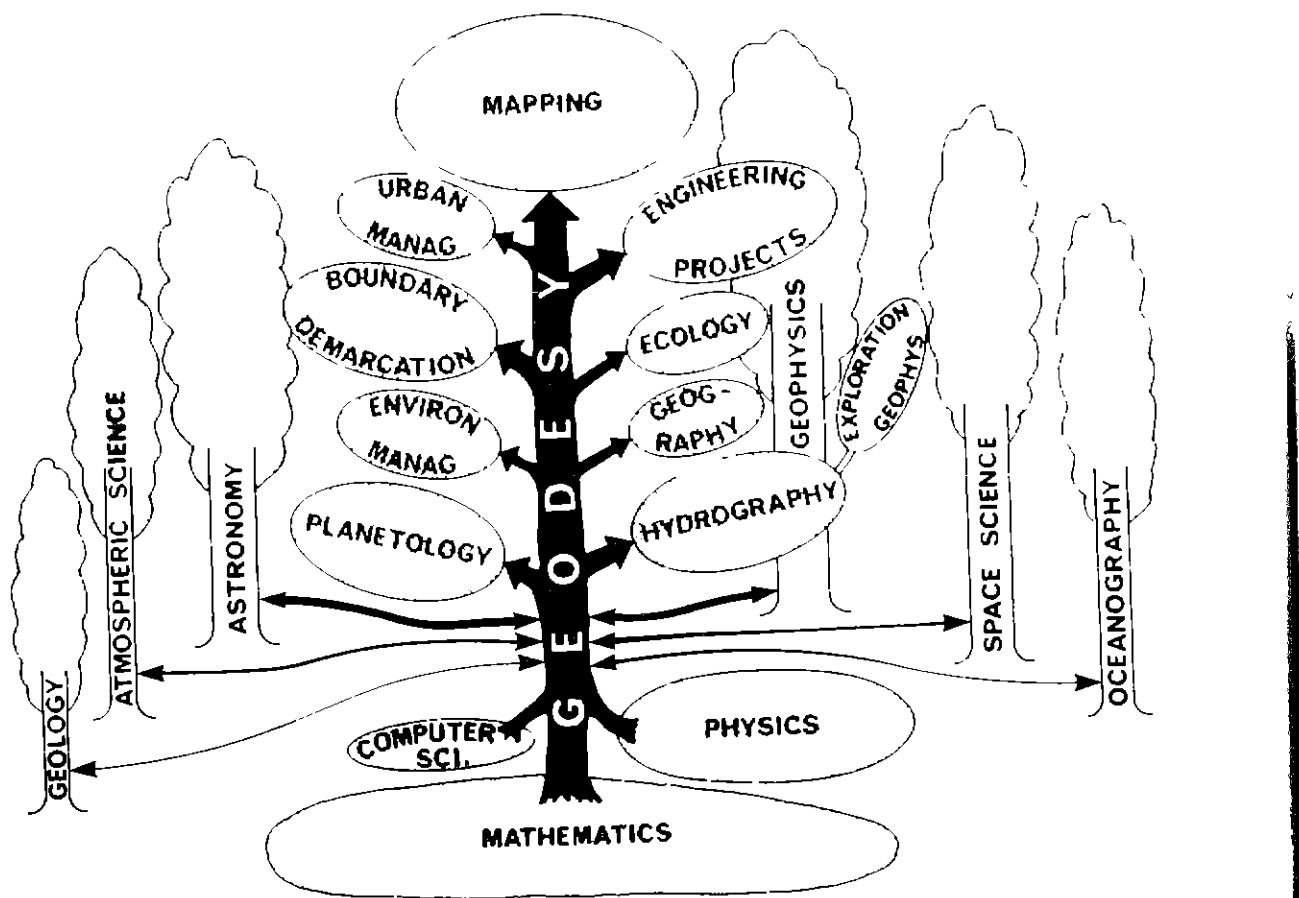


FIG. 2.1. Relation of geodesy to other disciplines.

of the collected data. As a significant number of these geodetic instruments use the visible part of the electromagnetic spectrum, the need for geometrical optics is clearly indicated.

Mechanics is required to understand the motions of the earth and its satellites (see Chapters 5 and 25). Two dynamical concepts are used in this context: the motion of a physical particle in a potential field (central as well as perturbed), and the rotation of a deformable body. Thus both the Keplerian and the perturbation [HAGIHARA, 1971] theories are needed together with the Liouville theory [ROUTH, 1884] of a deformable gyroscope. In this book, however, only Euler's theory (see §5.3) of a solid gyroscope is used; this theory is also needed in studying gyroscopic orientation (see §16.1).

Some elementary foundations of the mechanics of continuum and rheology aid in the appreciation of the earth's response to different stresses (see Chapters 8 and 25). Although an understanding of the physics of the earth's deformations is not a requirement in geodesy, an appreciation of these deformations is helpful. Further, rudiments of acoustics are used in marine positioning (see §18.4), and some knowledge of metrology is often employed in calibrating geodetic instruments.

A representation of all the relations described in this chapter is given in FIG. 1. Note the different shadings denoting the different relations.

## CHAPTER 3

# MATHEMATICS AND GEODESY

As stated in Chapter 2, geodesy has a special relationship with mathematics, and mathematics is extensively used in geodesy. The aim of this chapter is to describe the mathematics geodesists need. In pursuit of this goal though, we have decided to slant the presentation towards emphasizing the topics of particular importance in the subsequent chapters. In addition, some of the links between topics that the authors felt would be of assistance to the reader are shown, and any non-standard notation used in the text is defined. Most of the mathematics used herein can be found in KORN AND KORN [1968] and HOGG AND CRAIG [1970]. Nevertheless, other references are quoted in the text to offer a revealing perspective or a desired accent.

The review presented here is meant only to refresh the reader's memory and, as such, should not be used as a text on mathematics. Most, but not all, of these topics should commonly be found in an undergraduate mathematics curriculum of a university science or engineering course.

The topics are lumped together under four, more or less natural, headings: algebra, analysis, geometry, and statistics. No attempt is made to defend the didactic soundness of the presentation sequence or the appropriateness of the section selected for one topic or another. Topics that are very specialized, or considered clearly beyond the framework of an undergraduate mathematical curriculum, are treated directly in the main text of the book.

### 3.1. Algebra

In the first part of this section, that portion of mathematics dealing with the algebra of *vectors* and *matrices* is discussed. Throughout the book, vectors are denoted by lower case boldface letters ( $\mathbf{a}, \mathbf{b}, \dots$ ) and matrices by upper case boldface letters ( $\mathbf{A}, \mathbf{B}, \dots$ ). Unless stated otherwise, vectors will be considered as finite, ordered sequences of real numbers and thus can be regarded as belonging to real *vector*, or *linear*, *spaces*:  $\mathbf{a}$  that has  $n$  components ( $\dim(\mathbf{a}) = n$ ) is then

$$\mathbf{a} \in \mathbb{R}^n, \tag{3.1}$$

where  $\mathbb{R}$  denotes the *set of real numbers*. Similarly for matrices: if  $\dim(\mathbf{A}) = (n, m) = (\dim(\text{col } \mathbf{A}), \dim(\text{row } \mathbf{A}))$ , then  $\mathbf{A}$  belongs to space  $\mathbb{R}_{(n,m)}$ , isomorphic to  $\mathbb{R}^{nm}$ . Of

course, these vectors include the position vectors in geometrical spaces; nevertheless, the distinction is made in later chapters between these abstract vectors and vectors that carry obvious geometrical meaning (cf. §3.3).

Matrices (and for that matter vectors) can play one of the following two roles: they can be regarded either as objects or as *linear operators*. If we write

$$\mathbf{b} = \mathbf{B}\mathbf{a}, \quad (3.2)$$

then  $\mathbf{B}$  may be considered as transforming  $\mathbf{a} \in \mathcal{V}_a$  to  $\mathbf{b} \in \mathcal{V}_b$ , i.e., as a transformation operator between the two vector spaces  $\mathcal{V}_a, \mathcal{V}_b$ . The above equation presupposes *congruity* among  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $\mathbf{B}$ , i.e.,  $\dim(\mathbf{B}) = (\dim(\mathbf{b}), \dim(\mathbf{a}))$ . An example of a matrix usually regarded as an operator is the *Jacobian matrix* (of partial derivatives):

$$\frac{\partial \mathbf{f}}{\partial \mathbf{l}} = \begin{bmatrix} \frac{\partial f_1}{\partial l_1} & \frac{\partial f_1}{\partial l_2} & \cdots & \frac{\partial f_1}{\partial l_n} \\ \frac{\partial f_2}{\partial l_1} & \frac{\partial f_2}{\partial l_2} & \cdots & \frac{\partial f_2}{\partial l_n} \\ \vdots & & & \\ \frac{\partial f_m}{\partial l_1} & \frac{\partial f_m}{\partial l_2} & \cdots & \frac{\partial f_m}{\partial l_n} \end{bmatrix}, \quad (3.3)$$

where  $\mathbf{f} \in \mathcal{F}$  is a vector of functions  $f_i$  (a vector function of a vector of arguments  $l_j$ );  $\mathbf{l} \in \mathcal{L}$  is a *vector argument*; and  $\mathcal{F}, \mathcal{L}$  are vector spaces. Upon substitution of specific values, say  $l_1$ , for  $\mathbf{l}$ , one can view  $\partial \mathbf{f} / \partial \mathbf{l}$  as transforming the neighbourhood of  $l_1 \in \mathcal{L}$  onto a neighbourhood of  $\mathbf{f}(l_1) = \mathbf{f}_1 \in \mathcal{F}$ . An example of a matrix normally considered as an object in its own right is *Vandermonde's matrix*,

$$\Phi(\mathcal{T}) = \begin{bmatrix} \phi_1(\tau_1) & \phi_1(\tau_2) & \cdots & \phi_1(\tau_n) \\ \phi_2(\tau_1) & \phi_2(\tau_2) & \cdots & \phi_2(\tau_n) \\ \vdots & & & \\ \phi_u(\tau_1) & \phi_u(\tau_2) & \cdots & \phi_u(\tau_n) \end{bmatrix}, \quad (3.4)$$

composed of  $n$  functional values of  $u$  functions  $\phi_i$  of  $\tau_j \in \mathcal{T} \equiv \{\tau_1, \tau_2, \dots, \tau_n\}$ . These two roles are, of course, often interchanged.

Following are some particular kinds of matrices used in the text.

- (a) A *square matrix*  $Q$  is a matrix for which  $\dim(\text{col } Q) = \dim(\text{row } Q) = \dim(Q)$ .
- (b) A *symmetrical matrix*  $S$  is a square matrix for which  $s_{ij} = s_{ji}$ ,  $i, j = 1, \dots, \dim(S)$ , where  $s_{ij}$  are the *elements* of  $S$ .
- (c) An *antisymmetrical matrix*  $A$  is a square matrix for which  $a_{ij} = -a_{ji}$ ,  $i, j = 1, \dots, \dim(A)$ .
- (d) A *diagonal matrix*  $D$  is a square matrix for which  $d_{ij} = 0$ , for  $i \neq j$ . It is often denoted as  $D = \text{diag}(d_{ii}) = \text{diag}(d_i)$ . A matrix whose only non-zero elements are assembled in  $n$  diagonals spaced symmetrically around the main diagonal is called an  *$n$ -diagonal matrix*.

- (e) A *unit matrix*  $\mathbf{I}$  is a diagonal matrix for which  $d_i = 1$  for  $i = 1, \dots, \dim(\mathbf{I})$ .  
 (f) An *upper (lower) triangular matrix* is a square matrix for which all the elements below (above) the *main diagonal* are equal to zero.  
 (g) A *positive definite matrix*  $\mathbf{P}$  is a square matrix for which the *quadratic form*  $\mathbf{a}^T \mathbf{P} \mathbf{a}$  is a positive number for any  $\mathbf{a} \neq \mathbf{0}$ .  
 (h) A *dyadic matrix*  $\mathbf{Y}$  is a square matrix obtained as

$$\mathbf{Y} = \mathbf{a} \mathbf{a}^T, \quad (3.5)$$

where  $\mathbf{a}$  is an arbitrary vector; clearly,  $\dim(\mathbf{Y}) = (\dim(\mathbf{a}), \dim(\mathbf{a}))$ .

(i) An *orthogonal matrix* is a matrix  $\mathbf{R}$  for which all the columns (rows)  $\mathbf{c}_i$  satisfy the following equality:

$$\mathbf{c}_i^T \mathbf{c}_j = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \quad (3.6)$$

(j) If the elements in  $\mathbf{H}$  are themselves also matrices, then  $\mathbf{H}$  is called a *hypermatrix*. Any matrix  $\mathbf{B}$  whose larger dimension equals at least 2 becomes a hypermatrix by proper *partitioning*.

A *trace* of a square matrix  $\mathbf{Q}$  is a number defined as

$$\text{tr}(\mathbf{Q}) = \sum_{i=1}^{\dim(\mathbf{Q})} q_{ii}. \quad (3.7)$$

It can be shown that for a positive definite matrix  $\mathbf{P}$ , the following relation holds:

$$\mathbf{a}^T \mathbf{P} \mathbf{a} = \text{tr}(\mathbf{a} \mathbf{a}^T \mathbf{P}). \quad (3.8)$$

For any two matrices  $\mathbf{A}, \mathbf{B}$ ,

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}), \quad (3.9)$$

if both products exist.

Each square matrix  $\mathbf{Q}$  has a real number called a *determinant*,  $\det(\mathbf{Q})$ , associated with it. By definition

$$\det(\mathbf{Q}) = \sum_{(l)} \prod_{j=1}^{\dim(\mathbf{Q})} a_{jl}, \quad (3.10)$$

where the subscript  $l$ , assumes all possible values so that all permutations in which each column is represented only once are present and  $(l)$  is the number of interchanges necessary to bring the values of the second subscripts into natural order. The matrix  $\mathbf{Q}$  is called *regular* (non-singular) if and only if  $\det(\mathbf{Q}) \neq 0$ ; conversely, if  $\det(\mathbf{Q}) = 0$ , then  $\mathbf{Q}$  is *singular*. When  $\det(\mathbf{Q})$  is a very small number,  $\mathbf{Q}$  is known as *ill-conditioned*. A positive definite matrix is always regular; a diagonal matrix  $\mathbf{D}$  gives

$$\det(\mathbf{D}) = \prod_{i=1}^{\dim(\mathbf{D})} d_i. \quad (3.11)$$

The *rank* of a matrix  $\mathbf{B}$ ,  $\text{rank}(\mathbf{B})$ , is the number of linearly independent rows or columns of  $\mathbf{B}$ . If  $\mathbf{P}$  is positive definite, then  $\mathbf{B}\mathbf{P}\mathbf{B}^T$  is also positive definite when  $\text{rank}(\mathbf{B}) = \dim(\text{col } \mathbf{B}) \leq \dim(\text{row } \mathbf{B})$ . If for a square  $\mathbf{Q}$ ,  $\text{rank}(\mathbf{Q}) < \dim(\text{col } \mathbf{Q}) = \dim(\text{row } \mathbf{Q}) = \dim(\mathbf{Q})$ , then  $\mathbf{Q}$  is called *rank deficient*, with *defect*

$$\text{def}(\mathbf{Q}) = \dim(\mathbf{Q}) - \text{rank}(\mathbf{Q}). \quad (3.12)$$

A rank deficient matrix is singular [THOMPSON, 1969].

If a square matrix  $\mathbf{B}$  exists such that

$$\mathbf{B}\mathbf{Q} = \mathbf{I}, \quad (3.13)$$

it is called the *inverse* of  $\mathbf{Q}$  and is denoted by  $\mathbf{Q}^{-1}$ . Any regular (square) matrix has one and only one inverse. A singular matrix has no (regular) inverse as defined above; however, it may have other inverses known as singular or generalized, and these will be introduced in Part III. Some of the special kinds of matrices mentioned above have special (regular) inverses.

(a) The inverse of a symmetrical matrix is symmetrical.

(b) The inverse of a regular diagonal matrix  $\text{diag}(d_i)$  is

$$(\text{diag}(d_i))^{-1} = \text{diag}(d_i^{-1}). \quad (3.14)$$

(c) A square orthogonal matrix  $\mathbf{R}$  has an inverse

$$\mathbf{R}^{-1} = \mathbf{R}^T. \quad (3.15)$$

(d) For a hypermatrix  $\mathbf{H}$  one gets [FADDEEV AND FADDEEVA, 1963]

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{H}_{11}^{-1} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{B}_{11}^{-1} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \quad (3.16)$$

where either

$$\begin{aligned} \mathbf{B}_{22} &= (\mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12})^{-1}, & \mathbf{B}_{21} &= -\mathbf{B}_{22}\mathbf{H}_{21}\mathbf{H}_{11}^{-1}, \\ \mathbf{B}_{12} &= -\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{B}_{22}, & \mathbf{B}_{11} &= \mathbf{H}_{11}^{-1} - \mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{B}_{21}, \end{aligned} \quad (3.17)$$

if  $\mathbf{H}_{11}^{-1}$  and  $(\mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12})^{-1}$  exist, or

$$\begin{aligned} \mathbf{B}_{11} &= (\mathbf{H}_{11} - \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{21})^{-1}, & \mathbf{B}_{12} &= -\mathbf{B}_{11}\mathbf{H}_{12}\mathbf{H}_{22}^{-1}, \\ \mathbf{B}_{21} &= -\mathbf{H}_{22}^{-1}\mathbf{H}_{21}\mathbf{B}_{11}, & \mathbf{B}_{22} &= \mathbf{H}_{22}^{-1} - \mathbf{H}_{22}^{-1}\mathbf{H}_{21}\mathbf{B}_{12}, \end{aligned} \quad (3.18)$$

if  $\mathbf{H}_{22}^{-1}$  and  $(\mathbf{H}_{11} - \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{21})^{-1}$  exist. The immediate application of an inverse is in solving the *systems of simultaneous linear equations*. If

$$\mathbf{Bx} = \mathbf{l} \quad (3.19)$$

is such a system, then clearly

$$\mathbf{x} = \mathbf{B}^{-1}\mathbf{l} \quad (3.20)$$

is the solution, if the regular inverse  $\mathbf{B}^{-1}$  exists. Such systems and their solutions are investigated fully in Chapters 11 and 12. If they exist, the regular inverses have the following two often-quoted properties:

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}, \quad (k\mathbf{A})^{-1} = k^{-1}\mathbf{A}^{-1}. \quad (3.21)$$

Also the two matrix lemmas [MORRISON, 1969] which follow are useful:

$$(\mathbf{C}^{-1} + \mathbf{A}^T\mathbf{B}^{-1}\mathbf{A})^{-1} = \mathbf{C} - \mathbf{CA}^T(\mathbf{B} + \mathbf{ACA}^T)^{-1}\mathbf{AC}, \quad (3.22)$$

$$(\mathbf{C}^{-1} + \mathbf{A}^T\mathbf{B}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{B}^{-1} = \mathbf{CA}^T(\mathbf{B} + \mathbf{ACA}^T)^{-1}. \quad (3.23)$$

They are valid for an arbitrary  $\mathbf{A}$  and positive definite  $\mathbf{B}, \mathbf{C}$ .

Any symmetrical, positive definite matrix  $\mathbf{P}$  can be transformed to a diagonal form, i.e., diagonalized, in a variety of ways. One particular way, called *eigenvalue diagonalization*, is of special importance. It is based on the solutions  $(\lambda_1, \lambda_2, \dots, \lambda_{\dim(\mathbf{P})}) = \boldsymbol{\lambda}^T$  of

$$(\lambda \mathbf{I} - \mathbf{P}) = \mathbf{0}, \quad (3.24)$$

called *eigenvalues* of  $\mathbf{P}$ ; these are all positive. Vectors  $\mathbf{x}_i$  given by

$$(\lambda_i \mathbf{I} - \mathbf{P}) \mathbf{x}_i = \mathbf{0}, \quad i = 1, \dots, \dim(\mathbf{P}), \quad (3.25)$$

are called *eigenvectors* of  $\mathbf{P}$ . If the eigenvalues are all different, then there is one eigenvector for each eigenvalue  $\lambda_i$ ; all these eigenvectors are *mutually orthogonal*, i.e.,

$$\mathbf{x}_i \mathbf{x}_j^T = 0, \quad i \neq j. \quad (3.26)$$

If  $\mathbf{P}$  is real and symmetrical, then

$$\text{diag}(\lambda_i) = \mathbf{X}^{-1} \mathbf{P} \mathbf{X}, \quad (3.27)$$

where the columns of the matrix  $\mathbf{X}$  are the eigenvectors [THOMPSON, 1969]. The eigenvalue diagonalization can be interpreted as a transformation of  $\mathbf{P}$  into a coordinate system defined with its eigenvectors as a base.

It is of interest to know that for a symmetrical  $\mathbf{S}$ , the equation

$$\mathbf{x}^T \mathbf{S} \mathbf{x} = q = \text{const.} > 0, \quad (3.28)$$

where  $\mathbf{x}$  is taken as a position vector, is an equation of a *central quadric*. If, in addition,  $\mathbf{S}$  is positive definite, then the quadric is a *hyperellipsoid*; for  $\dim(\mathbf{x}) = 3$  we get an ellipsoid, and for  $\dim(\mathbf{x}) = 2$  an ellipse. The eigenvectors of  $\mathbf{S}$  give the directions of the main axes while  $2\lambda_i^{-1/2}$  are their lengths for  $q = 1$ .

A *complex number*, denoted throughout the book by an asterisk, is a special vector of two components;

$$z^* = (a, b), \quad (3.29)$$

where  $a$  is called the *real part* of  $z^*$ ;  $a = \text{re}(z^*)$ ; and  $b$  is the *imaginary part*,

$b = \operatorname{im}(z^*)$ . Written in the polar and exponential forms, it reads

$$z^* = A(\cos \psi + i \sin \psi) = A \exp(i\psi), \quad i = \sqrt{-1}, \quad (3.30)$$

where  $A$  is called the *amplitude*, or *absolute value*, of  $z^*$ ;  $A = |z^*|$ ; and  $\psi$  is the *argument* of  $z^*$ ,  $\psi = \arg(z^*)$ . These are given as

$$|z^*| = \sqrt{(a^2 + b^2)}, \quad \arg(z^*) = 2 \arctan\left(\frac{b}{A + a}\right) \pm 2k\pi, \quad k = 0, 1, 2, \dots \quad (3.31)$$

If  $a, b$  are real functions of  $x$ , then  $z^*$  is a *complex function* of  $x$ . The pair  $(a, -b) = A \exp(-i\psi) = A(\cos \psi - i \sin \psi)$  is called the *conjugate* of  $z^*$  and is denoted by  $\bar{z}^*$ . Using the conjugate, one gets

$$|z^*| = \sqrt{(z^* \times \bar{z}^*)}, \quad \arg(z^*) = (1/(2i)) \ln^*(z^*/\bar{z}^*), \quad (3.32)$$

where  $\ln^*$  denotes the main branch of a complex logarithm. Some of the properties of complex functions are:

- (a) If  $a = k(a_1 + a_2)$  and  $b = k(b_1 + b_2)$ , then  $z^* = (a, b) = k(z_1^* + z_2^*)$ , where  $z_1^* = (a_1, b_1)$ ,  $z_2^* = (a_2, b_2)$ .
- (b) If  $z^* = z_1^* \times z_2^*$ , then  $|z^*| = |z_1^*||z_2^*|$  and  $\arg(z^*) = \arg(z_1^*) + \arg(z_2^*) \pm 2k\pi$ ,  $k = 0, 1, 2, \dots$
- (c)  $\sin x = (\exp(ix) - \exp(-ix))/(2i)$  and  $\cos x = (\exp(ix) + \exp(-ix))/2$  [CHURCHILL AND BROWN, 1974].

### 3.2. Analysis

Analysis is that part of mathematics for which the notion of *limit* is the basis. Thus quantities from *locally compact spaces*, where it makes sense to define a *compact neighbourhood* of a point, will be considered here. *Sequences* and *series* of elements from locally compact spaces will also be used; the place of an element in a sequence or a series will be denoted here by a subscript. On the other hand, steps in iterative processes will be shown by superscripts in brackets.

The functions will always be considered *one valued*. They may be *real* (scalar), *vector*, or *matrix functions* of real (scalar) or vector arguments. One particular kind of function should be specifically mentioned: in the context of real functions, a function  $K$  that maps  $\mathbb{R}^n \times \mathbb{R}^n$  into  $\mathbb{R}$ , i.e.,

$$K \in \{\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}\}, \quad (3.33)$$

is known as a (real) *kernel*. For example, a distance,

$$\rho(P_i, P_j) \in \mathbb{R}, \quad (3.34)$$

between any two points  $P_i, P_j \in \mathbb{R}^n$  is clearly a kernel. If the relation between  $P_i, P_j$  described by the kernel  $K$  does not depend on the 'location' of  $P_i, P_j$  in  $\mathbb{R}^n$ , then one

speaks of a *homogeneous kernel*; if the relation does not depend on the ‘direction’ of  $P_i P_j$ , then one has an *isotropic kernel* [KREYSZIG, 1978]. Unless  $\mathcal{R}^n$  is a geometrical space, the two notions may be difficult to interpret.

Throughout the book,  $d$  will be used for a *differential*, with  $\delta$  or  $\Delta$  denoting small quantities. Derivatives of functions with respect to time (velocities) will be denoted by a dot over the functional symbol and second derivatives (accelerations) by two dots. The *total differential* of a *function of several (scalar) variables*  $f \in \{\mathcal{R}^n \rightarrow \mathcal{R}\}$  (whose *functional value* is  $f(x_1, x_2, \dots, x_n) = f(\mathbf{x})$ ) is

$$df = \sum_{i=1}^n \frac{\partial f}{\partial x_i} dx_i. \quad (3.35)$$

Here,  $\partial f / \partial x_i$  are the *partial derivatives* of  $f$  with respect to  $x_i$ . If the arguments are themselves functions of another independent variable, e.g.,

$$z = f(x_1(t), x_2(t), \dots, x_n(t)), \quad (3.36)$$

then one can also define the *total derivative* of  $f$  as follows:

$$\frac{df}{dt} = \sum_{i=1}^n \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}. \quad (3.37)$$

Note that, in particular, if  $z = f(x, y(x))$ , one has

$$\frac{df}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \frac{dy}{dx}. \quad (3.38)$$

Second and higher derivatives are defined similarly [PROTTER AND MORREY, 1973]. Often one deals with *functions of several vector variables*, e.g.,  $f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r)$ . The rules for the differentiation of such functions are exactly the same as if the arguments were scalars. Thus, for example, for symmetrical  $A$  we have

$$\frac{\partial}{\partial \mathbf{x}_1} (\mathbf{x}_1^T A \mathbf{x}_1 + \mathbf{B} \mathbf{x}_1 + \mathbf{C} \mathbf{x}_2) = 2 \mathbf{x}_1^T A + \mathbf{B}. \quad (3.39)$$

One of the standard problems connected with functions of a vector argument is the determination of a *local* or *absolute extremum*, i.e., either  $\min_{\mathbf{x} \in \mathcal{D}} f$  or  $\max_{\mathbf{x} \in \mathcal{D}} f$ . It is an awkward problem to solve under general circumstances. If  $f$  is known to have only one extremum of a kind in  $\mathcal{D} \subset \mathcal{R}^n$ , then the following system of equations,

$$\frac{\partial f}{\partial \mathbf{x}} = \mathbf{0}, \quad \mathbf{x} \in \mathcal{D}, \quad (3.40)$$

can be used in an attempt to get the solution  $\mathbf{x}_{\text{ext}}$ . Whether this is a minimum or maximum depends on the value of the second derivative [HANCOCK, 1917].

For vector functions, i.e.,  $f \in \{\mathcal{R}^n \rightarrow \mathcal{R}^n\}$ , one cannot use the ordinary rules for differentiation, and refuge has to be taken in *vector analysis* [WREDE, 1963; WILLIAMSON ET AL., 1972]. This branch of mathematics is based on the use of two

differential operators defined as follows:

$$\nabla = \left( \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \dots, \frac{\partial}{\partial x_n} \right) = \left( \frac{\partial}{\partial x_i}; i=1, \dots, n \right), \quad (3.41)$$

$$\Delta = \nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \dots + \frac{\partial^2}{\partial x_n^2} = \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}. \quad (3.42)$$

In three-dimensional curvilinear coordinates (see §3.3) these operators read

$$\nabla_q = \left( \frac{1}{H_i} \frac{\partial}{\partial q_i}; i=1, \dots, 3 \right), \quad (3.43)$$

$$\nabla_q^2 = \frac{1}{\prod_{i=1}^3 H_i} \sum_{j=1}^3 \frac{\partial}{\partial q_j} \left( \frac{\prod_{k=1}^3 H_k}{H_j^2} \frac{\partial}{\partial q_j} \right), \quad (3.44)$$

where  $H_i = dS_i/dq_i$  is Lamé's coefficient, and  $S_i$  is the length along the  $q_i$  coordinate line.

These operators can be applied to scalar or vector functions in a variety of ways. The  $\nabla$  operator will be applied here in four ways:

$$\begin{aligned} \nabla(a) &= \nabla a = \text{grad}(a) \text{ (vector)}; \\ &\text{through the scalar product } \nabla \cdot a = \text{div}(a) \text{ (scalar)}; \\ &\text{through the vector product } \nabla \times a = \text{curl}(a) \text{ (vector)}; \\ &\text{through the dyadic product } \nabla a^T \text{ (matrix)}; \end{aligned} \quad (3.45)$$

where  $a, a$  are scalar and vector functions of a vector argument in  $\mathbb{R}^3$ . The properties of  $\nabla$  are numerous and only a very few are reviewed here:

$$\begin{aligned} \nabla(ka) &= k \nabla(a) \text{ (linearity)}, \\ \nabla(a+b) &= \nabla(a) + \nabla(b) \text{ (linearity)}, \\ \nabla[a(b)] \cdot b &= \partial a / \partial |b| \text{ (projectivity)}, \\ \nabla \cdot r &= \text{div}(r) = 3 \text{ (divergence of radius vector)}, \\ \nabla r^T &= I, \end{aligned} \quad (3.46)$$

where  $a, b$ , and  $b$  are scalar and vector functions;  $r$  is a radius vector in  $\mathbb{R}^3$  (see §3.3); and  $k$  is a real number. Similar rules are valid for  $\nabla^2$ , from which only the following are excerpted here:

$$\nabla^2(ka) = k \nabla^2(a), \quad \nabla^2(a+b) = \nabla^2(a) + \nabla^2(b). \quad (3.47)$$

The development of a vector function of several vector variables into a *Taylor series* is a straightforward generalization of the same task in one dimension. It reads:

$$\begin{aligned} f(x_1, x_2, \dots, x_r) = & f(x_1^{(0)}, x_2^{(0)}, \dots, x_r^{(0)}) + \frac{\partial f}{\partial x_1} \Big|_{x_1=x_1^{(0)}} (x_1 - x_1^{(0)}) \\ & + \frac{\partial f}{\partial x_2} \Big|_{x_1=x_1^{(0)}} (x_2 - x_2^{(0)}) + \dots + \frac{\partial f}{\partial x_r} \Big|_{x_1=x_1^{(0)}} (x_r - x_r^{(0)}) \\ & \vdots \\ & + \text{higher-order terms.} \end{aligned} \quad (3.48)$$

Note that the partial derivatives are merely Jacobian matrices evaluated for the *points of expansion*  $x_1^{(0)}, x_2^{(0)}, \dots, x_r^{(0)}$  [KORN AND KORN, 1968]. While on the topic of development into power series, let us have a look at one particular scalar function: namely,

$$f(x; p) = (1 + p^2 - 2px)^{-1/2}, \quad (3.49)$$

where  $p$  is a parameter. Development into a (one-dimensional) *McLaurin series* (for  $x^{(0)}=0$ ) gives

$$(1 + p^2 - 2px)^{-1/2} = \sum_{n=0}^{\infty} p^n P_n(x), \quad (3.50)$$

where  $P_n(x)$  are known as *Legendre's functions*. The  $f(x; p)$  is called the *generating function* for the system of Legendre's functions. Many different generating functions and corresponding systems of functions are known [ABRAMOWITZ AND STEGUN, 1964]. These systems have a profound importance in solving differential equations, as will be seen later.

Turning now to integration, only integration in the Riemannian sense, i.e., the *Riemann integral*, is used in this book. From the existing special integrals, it is the *elliptical integrals* [REKTORYS, 1969],

$$y = \int_a^b (1 - k^2 \sin^2 x)^q dx, \quad (3.51)$$

that are mostly needed. Their evaluation is normally carried out by means of a power series. *Integrals of vector (matrix) functions* present no particular problem: they can be considered as vectors (matrices) of integrals and evaluated individually, component by component. A *convolution integral* is the name for an integral over a function *convolved* with a kernel, i.e., of a product of the function with the kernel. While one argument of the kernel becomes the *dummy variable*, the other becomes the argument of the result of the integration. The *mean value theorem* will also be needed in some parts of the book.

*Line integrals* are those that have a curve  $\mathcal{C}$  for the integration domain, and they are denoted as

$$y = \int_{\mathcal{C}} f d\mathcal{C}. \quad (3.52)$$

The problem with these integrals is that one must find a parameter that changes monotonically along  $\mathcal{C}$  to be used as the dummy variable. If  $\mathcal{C}$  is a *closed curve*, then the integral is written as

$$y = \oint_{\mathcal{C}} f d\mathcal{C}. \quad (3.53)$$

Similarly, *surface integrals* have a surface  $S$  for the integration domain. They are written as  $\iint_S f dS$ , or  $\iint_S f d\mathcal{S}$  if  $S$  is a *closed surface*. *Volume integrals* present a parallel situation. The equation that relates the surface and volume integrals is known as *Gauss's formula*, and it reads

$$\iint_S \mathbf{a} \cdot \mathbf{n} dS = \iiint_B \nabla \cdot \mathbf{a} dV. \quad (3.54)$$

Here  $\mathbf{a}$  is an arbitrary vector,  $S$  is the surface of  $B$ , and  $\mathbf{n}$  is a unit normal to  $S$ . Gauss's formula can also be written in terms of a scalar  $P$  instead of a vector  $\mathbf{a}$ . Let  $\mathbf{a} = \nabla P$ . Then, using the projectivity of the  $\nabla$  operator (eqn. (46)), we get

$$\mathbf{a} \cdot \mathbf{n} = \nabla P \cdot \mathbf{n} = \frac{\partial P}{\partial n}, \quad (3.55)$$

and substitution into Gauss's formula yields

$$\iint_S \frac{\partial P}{\partial n} dS = \iiint_B \nabla^2 P dV. \quad (3.56)$$

Consideration of two arbitrary scalar functions  $P$  and  $Q$ , applications of Gauss's formula to  $\mathbf{a} = P \nabla Q$  and  $\mathbf{b} = Q \nabla P$ , and subtraction of the second from the first yield *Green's second identity*:

$$\iint_S \left( \frac{P \partial Q}{\partial n} - \frac{Q \partial P}{\partial n} \right) dS = \iiint_B (P \nabla^2 Q - Q \nabla^2 P) dV \quad (3.57)$$

(valid for any two scalar functions  $P$  and  $Q$ ).

Often it is required to change the dummy variables in the integrals. This is done through the following equation:

$$\int_{V(x)} \cdots \int f(x) dx = \int_{V(x^{-1})} \cdots \int f(x(y)) \left| \frac{\partial x}{\partial y} \right| dy = \int_{V(y)} \cdots \int g(y) \left| \frac{\partial x}{\partial y} \right| dy, \quad (3.58)$$

where  $| \partial x / \partial y | = \det(\partial x / \partial y)$  is the determinant of the Jacobian matrix, called simply the *Jacobian* [FLANDERS AND PRICE, 1978];  $V$  is the volume over which the integration is to be performed; and  $x^{-1}$  is the inverse function to  $x$ , i.e.,  $y = x^{-1}(x)$ .

Very often in the physical sciences, the form of the functional relation between two quantities, say  $x$  and  $y$ , is not known and cannot even be guessed. On the other hand, the relation between the differentials  $dx$  and  $dy$  can be formulated; this relation is called an *ordinary differential equation* of  $m$ th order, depending on the

highest ( $m$ th) power of the differentials, i.e., the highest derivative in the relation. Of these, ordinary differential equations of second order are the most important. Once the ordinary differential equation of  $m$ th order is formulated, the functional relation, say  $y = y(x)$ , can be sought for  $x \in \mathfrak{D} \equiv \langle a, b \rangle \subset \mathfrak{R}$ . It contains as many undetermined parameters, called *integration constants*, as is the order of the ordinary differential equation. To obtain a unique solution, called a *particular solution*,  $m - 1$  values of  $y$  (or its derivatives) must be known on  $\langle a, b \rangle$ . These are usually either the *initial values*  $y(a), y'(a), \dots$  or the *boundary values*  $y(a), y(b), \dots$ . One speaks of either an *initial value problem* or a *boundary value problem*. If the function being sought is a vector function of a scalar argument, the *vector differential equation* can be written as a *system of ordinary differential equations* for each component.

One particular boundary value problem must be mentioned here: this is *Sturm–Liouville's boundary value problem*. It is based on the following ordinary differential equation:

$$(Ky')' + (\lambda\rho - q)y = 0, \quad (3.59)$$

where  $K(x) > 0, \rho(x) \in \langle 0, k \rangle, q(x) \geq 0$  on  $\langle a, b \rangle$  (and  $K, K', q, \rho$  all continuous on  $\langle a, b \rangle$ ) are given while  $\lambda \in \mathfrak{R}$  is not specified. The boundary values are prescribed as

$$y(a) = y(b) = 0 \quad (3.60)$$

This boundary value problem has particular solutions only for particular values of  $\lambda \geq 0$ , called *eigenvalues* of the problem. The particular solutions corresponding to the individual eigenvalues are called the *eigenfunctions* of the problem: there are always infinitely many eigenvalues and eigenfunctions. The *complete solution* to the Sturm–Liouville problem is then given as a linear combination of these infinitely many eigenfunctions. Different choices of  $K, \rho, q, a, b$  lead to different systems of eigenvalues and eigenfunctions. The Legendre functions seen above are one example, stemming from *Legendre's equation*, a special case of Sturm–Liouville's boundary value problem for  $K(x) = 1 - x^2, \rho(x) = 1, q(x) = 0, \mathfrak{D} \equiv \langle -1, 1 \rangle$ . For  $K(x) = \rho(x) = 1, q(x) = 0$ , the equation of *simple harmonic motion* is obtained. Specification of  $\mathfrak{D} \equiv (-\pi, \pi)$  leads to the system of eigenvalues  $\lambda = 0, 1, 4, 9, 16, \dots$  and eigenfunctions  $\{\cos \sqrt{\lambda} x, \sin \sqrt{\lambda} x; \lambda = 0, 1, 4, \dots\}$  [GREENBERG, 1971].

All the systems of eigenfunctions  $\{\phi_n; n = 1, 2, \dots\}$  are orthogonal with weight  $W$ . This means that the *scalar*, also called *inner product* of  $\phi_n$  and  $\phi_s$  is equal to zero if  $n \neq s$  or, more formally,

$$\int_{\mathfrak{D}} W(x) \phi_n(x) \phi_s(x) dx = \begin{cases} 0, & n \neq s, \\ \|\phi_n\|^2, & n = s, \end{cases} \quad (3.61)$$

where the expression  $\|\phi_n\| = \sqrt{\int_{\mathfrak{D}} W(x) \phi_n^2(x) dx}$  is called the *norm* of  $\phi_n$ . Any function  $f \in \{\mathfrak{D} \rightarrow \mathfrak{R}\}$  can be developed into an *eigenfunction series* as follows:

$$f(x) = \sum_{n=0}^{\infty} c_n \phi_n(x), \quad (3.62)$$

where

$$c_n = \|\phi_n\|^{-1} \int_{\mathcal{B}} W(x) \phi_n(x) f(x) dx. \quad (3.63)$$

An example of such a development is *Fourier's trigonometrical series* that uses the trigonometrical eigenfunctions shown above. The eigenfunction series are sometimes also called the *generalized Fourier series*. Clearly, the power series (50) is also an eigenfunction series representing the generating function, given by (49), with  $c_n = p^n$ .

When one deals with functions of several independent variables, the differential formulation leads to *partial differential equations*. In formulating the partial differential equations, naturally, heavy use is made of vector analysis. For instance, KOCHIN [1961] has shown that a *vector field*  $f \in \{\mathbb{R} \rightarrow \mathbb{R}^3\}$  can be fully described once its divergence and curl are known. These two differential operations can be understood, in this context, as four partial differential equations of first order for  $f$ . As another example,

$$\nabla^2 f(\mathbf{r}) = g(\mathbf{r}) \quad (3.64)$$

is a second-order, *unhomogeneous partial differential equation* for the *scalar field*  $f \in \{\mathbb{R}^n \rightarrow \mathbb{R}\}$ , called *Poisson's equation*. When  $g(\mathbf{r}) = 0$ , the *homogeneous equation* is known as *Laplace's equation*. These are the two main partial differential equations used in geodesy. A solution of Laplace's equation is known as a *harmonic function*; it has many useful properties [MACMILLAN, 1930].

As in the case of the ordinary differential equations, when a unique solution is required, then values of the desired solution  $f$  must be known on the boundary  $\mathcal{S}$  of the domain  $\mathcal{B}$  of the partial differential equation; one again speaks of a boundary value problem. In the case of Laplace's equation, it is called *Dirichlet's boundary value problem*. It has been shown to have a unique solution  $f$ , if  $\mathcal{S}$  is smooth enough [LANDKOF, 1972]. In some cases, the values of  $f$  on  $\mathcal{S}$  are not available. Instead, one may be able to specify the values of the normal derivative of  $f$  (the derivative with respect to the normal to  $\mathcal{S}$ ): this is known as *Neumann's boundary value problem*. At other times, none of these boundary values are known, but their linear combination may be obtained on  $\mathcal{S}$ : this is a boundary value problem of a mixed type.

Ways of solving a boundary value problem in  $n$  dimensions are many; only three, however, are ordinarily used in geodesy. The first—the *method of separation of variables*—is based on the idea of expressing the function being sought as a product of  $n$  independent functions of one variable. Such a substitution transforms the partial differential equation into a system of  $n$  ordinary differential equations. If these ordinary differential equations happen to be of a Sturm–Liouville type, then the product of the systems of their eigenfunctions is the system of eigenfunctions (of several variables) for the partial differential equation. The solution of the boundary value problem is then obtained as that linear combination of the eigenfunctions that satisfies the boundary values. But that particular linear combination is merely the development of the boundary values into the eigenfunction series (of several variables) as seen earlier (for one dimension), and one speaks about *Fourier's method*

[WYLIE, 1966]. The second method is known as *Green's method*. Its idea is to find such a kernel, called *Green's function*, which, convolved with the boundary values on  $\mathbb{S}$ , gives the solution (cf. the convolution integral). The form of Green's function depends on the shape of  $\mathbb{S}$ , and to find it may not be simple [GREENBERG, 1971]. The last technique is based on the use of Gauss's formula (54), by means of which the boundary value problem may be transformed to an *integral equation* [HOHEISEL AND TROPPER, 1963]. In this book, the results of such transformations will be *Fredholm's linear integral equations* of the following type [JASWON AND SYMM, 1977]:

$$f(\mathbf{r}) + k \int_{\mathbb{S}} K(\mathbf{r}, \mathbf{r}') f(\mathbf{r}') d\mathbb{S} = g(\mathbf{r}). \quad (3.65)$$

These are usually solved by approximating the compact integration domain by its discrete approximation, thus converting the problem to one of solving a system of simultaneous linear algebraic equations (see §3.1).

### 3.3. Geometry

Under this somewhat general heading, two broader topics of mathematics, pertinent to geodesy, will be discussed: a geometrical treatment of coordinate systems of various kinds, and differential geometry. From the methodological point of view, geodesy is simply applied geometry, thus the geodesist really needs all the tools geometry can offer. The above were selected only because they are the most widely used.

When speaking about a geometrical approach to coordinate systems and their transformations, it is clearly helpful to begin with the idea of three-dimensional *geometrical spaces*  $\mathbb{E}_3$  in the classical Euclidean sense. In such spaces it is possible, and usually desirable, to define a *Cartesian coordinate system*, i.e., an orthonormal, system (with the same scale on all three axes) with *coordinates* taken as segments on coordinate axes denoted by  $(x, y, z)$  or  $(x_1, x_2, x_3)$ . A position of a *point*  $P$  in  $\mathbb{E}_3$ , using the Cartesian system, is specified by assigning real numbers to  $x$ ,  $y$ , and  $z$ . Such a triplet of numbers is called the *position vector*, or *radius vector*, of  $P$  and is denoted by  $\bar{r} = (x, y, z)$ . Thus  $\bar{r}$  also denotes a point in  $\mathbb{E}_3$ ; clearly,  $\bar{r}$  is only a special case of  $\mathbf{r}$  from §3.1 and §3.2. The *distance* between two points  $\bar{r}_1, \bar{r}_2$  in  $\mathbb{E}_3$  is measured by the usual Euclidean metric:

$$\rho(\bar{r}_1, \bar{r}_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}. \quad (3.66)$$

Coordinates  $\bar{r}$  in one Cartesian system can be transformed to another Cartesian system, with different *origin* and different *orientation* but the same *polarity*, through the following *transformation*:

$$\bar{r}' = \mathbf{R}(\omega_1, \omega_2, \omega_3) \bar{r} + \bar{r}'_0. \quad (3.67)$$

Here,  $\bar{r}'_0$  is the position vector of the origin of the first in the second system and is

called the *translation vector*. The rotation matrix  $\mathbf{R}$  rotates the first system into the second around the axes of the first by the angles  $\omega_1, \omega_2, \omega_3$ . It is often written as

$$\mathbf{R}(\omega_1, \omega_2, \omega_3) = \mathbf{R}_1(\omega_1)\mathbf{R}_2(\omega_2)\mathbf{R}_3(\omega_3), \quad (3.68)$$

where

$$\begin{aligned} \mathbf{R}_1(\omega_1) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \omega_1 & \sin \omega_1 \\ 0 & -\sin \omega_1 & \cos \omega_1 \end{bmatrix}, & \mathbf{R}_2(\omega_2) &= \begin{bmatrix} \cos \omega_2 & 0 & -\sin \omega_2 \\ 0 & 1 & 0 \\ \sin \omega_2 & 0 & \cos \omega_2 \end{bmatrix}, \\ \mathbf{R}_3(\omega_3) &= \begin{bmatrix} \cos \omega_3 & \sin \omega_3 & 0 \\ -\sin \omega_3 & \cos \omega_3 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \end{aligned} \quad (3.69)$$

are the *fundamental rotation matrices* describing rotations around  $x$ ,  $y$ , and  $z$  axes respectively. These being orthogonal matrices, we have  $\mathbf{R}_i^{-1}(\omega_i) = \mathbf{R}_i^T(\omega_i) = \mathbf{R}_i(-\omega_i)$ , and the inverse transformation to (67) reads:

$$\bar{r} = \mathbf{R}_3(-\omega_3)\mathbf{R}_2(-\omega_2)\mathbf{R}_1(-\omega_1)(\bar{r}' - \bar{r}'_0). \quad (3.70)$$

If the polarity of the two systems is different, then the *reflection matrix*  $\mathbf{P}_2$ ,

$$\mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (3.71)$$

is used to premultiply  $\bar{r}$  and  $\bar{r}' - \bar{r}'_0$  respectively [THOMPSON, 1969]. It should be noted that rotation matrices are not commutative. It is, however, easy to chain them together when several systems are involved in the transformation; they are associative and can be graphically displayed in *commutative diagrams*. Rotation matrices are also a natural instrument to use in solving problems from *spherical trigonometry*.

*Curvilinear coordinates*  $(q_1, q_2, q_3)$  are derived from Cartesian through

$$q_i = q_i(x, y, z), \quad i = 1, 2, 3. \quad (3.72)$$

Inverse formulae are also required to be valid. Equations (72) may have none, one, or several parameters. An example of a *non-parametric coordinate system* is the system of *spherical coordinates*; a one-parametric *ellipsoidal system* — with the parameter being the focal length  $E$  — is shown in FIG. 20.4, and a two-parametric *geodetic ellipsoidal system* — with parameters being the size and shape of the ellipsoid the system uses, *geodetic latitude*  $\phi$  and *geodetic longitude*  $\lambda$  being the coordinates — is shown in FIG. 7.4. If several curvilinear systems have a common origin and direction of axes, then these systems are known as a *family of coordinate systems*. To each family there belongs one representative Cartesian system; if a transformation from one curvilinear system into a curvilinear system from another family is needed, it goes through their respective representative Cartesian systems by means of (67) and (70).

For transformations within one family, eqns. (72) and their inverses are used. These can be linearized using a Taylor series, if the validity of the transformation formulae can be limited to a neighbourhood of the point of expansion (cf. §3.2). Then, the transformation  $\bar{x} \rightarrow \bar{q}$  becomes *locally linear* and is once more mediated by the Jacobian matrix. If the Jacobian matrix has orthogonal columns, the  $q$  system is *locally orthonormal*. In this case, the inverse linear transformation is done through the transpose of the Jacobian matrix. The above discussed rotation, reflection, and translation are globally orthogonal transformations, and the first two are also linear.

A *spatial curve* can be described in a variety of ways, the simplest being

$$\bar{r} = \bar{r}(S) = (x(S), y(S), z(S)), \quad (3.73)$$

where  $S$  is a scalar parameter on the curve, typically the *arc length*. Often, the curve can only be formulated in terms of a vector differential equation, such as

$$f_1(\bar{r}) \frac{d^2\bar{r}}{dS^2} + f_2(\bar{r}) \frac{d\bar{r}}{dS} = f_3(\bar{r}), \quad (3.74)$$

or as differential equations of projections of  $\bar{r}(S)$  onto coordinate planes, i.e.,

$$\frac{dx}{f_x(\bar{r})} = \frac{dy}{f_y(\bar{r})} = \frac{dz}{f_z(\bar{r})} \quad \text{for } f_x(\bar{r}) \cdot f_y(\bar{r}) \cdot f_z(\bar{r}) \neq 0. \quad (3.75)$$

In the latter case, each equation gives a plane curve in two dimensions, e.g.,  $y = y(x)$ , which is then interpreted as a cylinder in three dimensions. A similar possibility is to have the curve defined as an intersection of two surfaces,  $z = z_1(x, y)$  and  $z = z_2(x, y)$ —see eqn. (81).

Let us mention here that a spatial curve possesses both a *curvature* and a *torsion*. The curvature takes place in the *osculating plane* defined by three infinitesimally close points on the curve. The torsion is defined as the curvature in the *rectifying plane*, i.e., in the plane perpendicular to both the osculating and the *normal planes*. The curvature  $\kappa$  and torsion  $\tau$  are related through *Frenet's formulae*:

$$\frac{d\bar{t}}{dS} = \kappa\bar{n}, \quad \frac{d\bar{n}}{dS} = -\kappa\bar{t} + \tau\bar{b}, \quad \frac{d\bar{b}}{dS} = -\tau\bar{n}, \quad (3.76)$$

where  $\bar{t}, \bar{n}, \bar{b}$  are *tangent*, *normal*, and *binormal* unit vectors [PROTTER AND MORREY, 1973]. Since

$$\bar{t} = \frac{d\bar{r}}{dS}, \quad (3.77)$$

the curvature is related to the second derivative of  $\bar{r}$ . This is seen even more clearly in the expression for the curvature of a plane curve  $y = y(x)$ :

$$\kappa = y''(1 + (y')^2)^{-3/2}. \quad (3.78)$$

The *radius of curvature* is the inverse of  $\kappa$  and can also be written as

$$R = \kappa^{-1} = \frac{dS}{d\alpha}, \quad (3.79)$$

where  $d\alpha$  is the change in the direction of the normal vector, corresponding to  $dS$ .

A *surface* in three-dimensional space can again be given by one of several possible ways. Most often in geodesy, the implicit formula

$$W(\vec{r}) = W(x, y, z) = 0 \quad (3.80)$$

is used. The explicit expression

$$z = z(x, y) \quad (3.81)$$

is seldom available except when the surface is given by a generalized two-dimensional polynomial (cf. eqn. (62))

$$z(x, y) = \sum_{n=0}^N c_n \phi_n(x, y) = \tilde{\Phi}^T(x, y) \mathbf{c}, \quad (3.82)$$

where  $\tilde{\Phi}(x, y)$  is one column of Vandermonde's matrix (cf. eqn. (4)). Let us mention in passing that an explicit expression (81) for a plane can always be written in a linear form, i.e.,

$$z(x, y) = \frac{\partial z}{\partial x} x + \frac{\partial z}{\partial y} y + \text{const.} \quad (3.83)$$

Very popular in geodesy is the use of *coordinate surfaces*, i.e.,

$$q_i = \text{const.}, \quad (3.84)$$

called *datums* for various tasks. One such datum may be a *sphere* ( $r = a$ ) in a spherical coordinate system, or the *ellipsoid* ( $h = 0$ ) in the above mentioned (ellipsoidal) geodetic system.

Generally, a surface has at a point a curvature different in different directions. A normal plane, containing the normal to the surface at  $P$ , is the osculating plane of the normal section made by this plane. It is the curvature of this normal section that defines the curvature of the surface at  $P$  in the direction of the normal plane. This relation is, however, generally valid only at  $P$ ; the curve for which this relation is valid everywhere is the *geodesic curve*, also called simply the *geodesic*. The geodesic has no curvature in the tangent plane; it is locally straight on the surface. The geodesic also has one more outstanding property: it is that unique curve  $\tilde{\mathcal{C}}$  having the shortest possible length, that connects any two points on an ordinary surface  $S$ . One has, by definition,

$$\min \int_{\mathcal{C}} dS \Rightarrow \tilde{\mathcal{C}}, \quad (3.85)$$

where  $\int_{\mathcal{C}} dS$  (cf. eqn. (52)) is, of course, the length of the curve  $\mathcal{C}$  between the two fixed points on  $S$  [SYNGE AND SCHILD, 1949].

Let us take the plane tangent to  $\mathbb{S}$  at  $P$  and plot the lengths of the radii of curvature in the corresponding directions. At any *regular point* on  $\mathbb{S}$ , the pattern obtained would generally be an ellipse; in some cases it will change to a circle, hyperbolas or a pair of parallel lines (for points on a linear surface). This pattern is known as *Dupin's indicatrix*, and it serves to distinguish *elliptical*, *circular*, and *hyperbolic* points on  $\mathbb{S}$ . Its shape is identical with that one obtains by cutting  $\mathbb{S}$  with a plane parallel with but differentially displaced from the tangent plane. It is a close relative of *Tissot's indicatrix*, which is used heavily in the theory of mappings (cf. §16.3). Tissot's indicatrix is given through a quadratic form (cf. eqn. (28))

$$\mathbf{u}^T \mathbf{G} \mathbf{u} = \text{const.}, \quad (3.86)$$

where  $\mathbf{u} = (u_1, u_2)^T$  is a two-dimensional vector in the tangent plane whose components are given by the parameters on  $\mathbb{S}$ ,  $u_1$ , and  $u_2$ . The shape of the indicatrix is clearly dictated by the matrix  $\mathbf{G}$ ,

$$\mathbf{G} = \begin{bmatrix} e & f \\ f & g \end{bmatrix}, \quad (3.87)$$

composed of the *Gaussian fundamental quantities*. These are evaluated from

$$e = \left| \frac{\partial \vec{r}}{\partial u_1} \right|^2, \quad f = \frac{\partial \vec{r}}{\partial u_1} \cdot \frac{\partial \vec{r}}{\partial u_2}, \quad g = \left| \frac{\partial \vec{r}}{\partial u_2} \right|^2, \quad (3.88)$$

where  $\vec{r} = \vec{r}(u_1, u_2)$  is the equation of  $\mathbb{S}$  in terms of the two parameters  $u_1, u_2$ .

Denoting the maximum and minimum radii of curvature by  $R_{\min}, R_{\max}$ —these occur in the directions  $\alpha_{\min}, \alpha_{\max}$  of the eigenvectors of Dupin's indicatrix—the radius of curvature in the direction  $\alpha$  is obtained from *Euler's formula* [MCCONNELL, 1931]

$$\frac{1}{R_\alpha} = \pm \frac{\cos^2 \alpha}{R_{\min}} + \frac{\sin^2 \alpha}{R_{\max}}. \quad (3.89)$$

If the radius of curvature of a normal section is  $R$ , then the radius of curvature  $R(\theta)$  of a section cut by a plane, inclined at an angle  $\theta$  with respect to the normal plane, is

$$R(\theta) = R \cos \theta \quad (3.90)$$

(Meusnier's theorem).

### 3.4. Statistics

A sequence of  $N$  real numbers  $l_i$ , usually obtained as a result of some measurements, is called a *random sample*. These numbers can be put together into groups according to their magnitudes, and a *histogram* or *polygon* can be plotted (see FIG.

13.1). The *sample mean* is given as

$$\bar{l} = \frac{1}{N} \sum_{i=1}^N l_i = \sum_{j=1}^M \text{pr}_j \tilde{l}_j, \quad (3.91)$$

and the *sample variance* as

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (l_i - \bar{l})^2 = \sum_{j=1}^M \text{pr}_j (\tilde{l}_j - \bar{l})^2, \quad (3.92)$$

where  $\text{pr}_j = n_j/N$ ,  $j = 1, \dots, M \leq N$  are the *relative counts* of the elements in the  $M$  groups, and  $\tilde{l}_j$  are the group values [WONNACOTT AND WONNACOTT, 1972]. The probability  $\text{pr}_j$  can also be regarded as the *experimental probability* that  $l_i$  will be in the  $j$ th group. Samples with smaller  $s^2$ , and thus smaller *standard deviation*  $s$ , will be referred to as more accurate. The distinction between *accuracy* and *precision* is not going to be made, it being a matter of *bias* which will only be tackled in §13.1. The quantity  $s/l$  is referred to as the *relative accuracy*.

The *stochastical (random) variable*  $l$  is a variable from  $\mathcal{R}$  such that with each value of  $l \in \mathcal{R}$  there is associated a non-negative real number  $\phi(l)$ . The real function  $\phi \in \{\mathcal{R} \rightarrow [0, \infty)\}$  is called the *probability density function* of  $l$ . Normally, a probability density function contains several *distribution parameters*  $\theta_1^{(l)}, \theta_2^{(l)}, \dots \in \mathcal{R}$  and one speaks about one-, two-, or multi-parametric probability density functions. This function will be coded either as  $\phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)}, \dots)$  or  $\phi_l(\xi), \phi_l$ , or even  $\phi(l)$  when there is no danger of confusion. *Normal*, *uniform*,  $\chi^2$  (*chi-squared*) probability density functions are some well-known examples. A function of a random variable is itself a random variable with generally a different probability density function. Real numbers

$$\mu_l = \int_{\mathcal{R}} \phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)}, \dots) \xi d\xi, \quad (3.93)$$

$$\sigma_l^2 = \int_{\mathcal{R}} \phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)}, \dots) (\xi - \mu_l)^2 d\xi, \quad (3.94)$$

are called the *mean of the probability density function* and the *variance of the probability density function*. Clearly, there is an affinity between (91) and (93) as well as between (92) and (94). The probability density function is used to assign *probability* (values) to *probability statements*, e.g., what is the probability that  $l$  is between  $a$  and  $b$ ? The answer is

$$\text{pr}(a < l < b) = \int_a^b \phi_l(\xi) d\xi. \quad (3.95)$$

From this equation, one can see that

$$\int_{\mathcal{R}} \phi_l(\xi) d\xi = 1 \quad (3.96)$$

must be satisfied for any probability density function.

*Stochastical (random) multivariate*  $l \in \mathcal{R}^n$  is a direct generalization of the random variable (univariate) into  $n$  dimensions. Its probability density function,  $\phi \in \{\mathcal{R}^n \rightarrow$

$\langle 0, \infty \rangle$ , is a function of  $n$  variables and can be written as  $\phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)}, \dots)$ . The distribution parameters are themselves multidimensional quantities: The *mean of the multidimensional probability density function of  $l$*  is given as (cf. eqn. (93))

$$\mu_l = \int_{\mathcal{R}^n} \phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)}, \dots) \xi d\xi. \quad (3.97)$$

Taking two components  $l_i, l_j$  of  $l$ , one can speak of their *statistical dependence* or *statistical independence* [WONNACOTT AND WONNACOTT, 1972]. Statistical dependence of various degrees is an intrinsic property of  $l$  and has its roots in the nature of, or mode of acquisition of,  $l$ . A deeper inquiry into the nature of statistical dependence is included in §10.3.

The *expectation operator* is an operator that acts upon any function of a random multivariate (that includes the univariate as well) as follows:

$$E(f(l)) = \int_{\mathcal{R}^n} \phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)}, \dots) f(\xi) d\xi \in \mathcal{R}^m. \quad (3.98)$$

Note that the dimensionality ( $m$ ) of the function does not have to correspond to the dimensionality ( $n$ ) of the multivariate. The role of  $\phi_l$  in the expectation operator is that of a parameter function. Using the expectation operator, (93), (94) and (97) can be rewritten as

$$\mu_l = E(l), \quad \sigma_l^2 = E[(l - \mu_l)^2], \quad \mu_l = E(l). \quad (3.99)$$

The expectation of the dyadic matrix  $(l - \mu_l)(l - \mu_l)^T$ , i.e.,

$$C_l = E[(l - \mu_l)(l - \mu_l)^T], \quad (3.100)$$

is called the *covariance matrix* of  $l$  and is probably the most important quantity in the multivariate statistics used in geodesy. Denoting (cf. eqn. (99))

$$\sigma_i^2 = E[(l_i - \mu_{l,i})^2] \quad (3.101)$$

and, analogously

$$\sigma_{ij} = \sigma_{ji} = E[(l_i - \mu_{l,i})(l_j - \mu_{l,j})], \quad (3.102)$$

the covariance matrix can be written as

$$C_l = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix}. \quad (3.103)$$

It is natural to call the quantities defined by (101) variances of the components of  $l$  and those defined by (102) covariances. Covariance  $\sigma_{ij}$  is a certain measure of statistical dependence between  $l_i$  and  $l_j$ . If  $l_i$  and  $l_j$  are statistically independent, then

$\sigma_{ij} = 0$ ; in the opposite case, they are usually, but not always, statistically dependent [HOGG AND CRAIG, 1970]. Very often in practice, the degree of statistical dependence of two components  $l_i, l_j$  of a random multivariate is not known and is to be estimated from the appropriate samples (see §13.1). This is known as the determination of the degree of correlation. That technique is not used anywhere in the book; the term correlation coefficient refers here to the expression

$$\rho_{ij} = \sigma_{ij}/(\sigma_i \sigma_j). \quad (3.104)$$

The  $E$  operator is linear in the sense of the individual components of  $l$ , i.e.,

$$E(l_i + l_j) = E(l_i) + E(l_j). \quad (3.105)$$

It is also commutative with respect to a multiplication by a non-probabilistic matrix (and thus even by a vector and a scalar):

$$E[Kf(l)] = KE[f(l)], \quad (3.106)$$

and

$$E(K) = K. \quad (3.107)$$

Algebraic rules for probability statements, similar to those for expectations, exist [FREUND, 1971]. Denoting by  $e_i$  events for which probabilities are sought, e.g.,  $e \equiv a < l < b$  in (95), then, for disjoint  $e_i$ , we have

$$\text{pr} \bigcup_i e_i = \sum_i \text{pr}(e_i). \quad (3.108)$$

If the events are statistically independent, then the simultaneous probability is

$$\text{pr} \bigcap_i e_i = \prod_i \text{pr}(e_i). \quad (3.109)$$

The conditional probability  $\text{pr}(e_i/e_j)$  of  $e_i$  occurring, given the knowledge that  $e_j$  has already occurred, is

$$\text{pr}(e_i/e_j) = \text{pr}(e_i \text{ and } e_j)/\text{pr}(e_j), \quad (3.110)$$

and makes sense only if  $\text{pr}(e_j) \neq 0$ .

## CHAPTER 4

# STRUCTURE OF GEODESY

In the previous three chapters we have talked a great deal about geodesy but, up to this point, we have neither strictly defined it nor even directly delineated its scope. Section one is devoted to just this task. The difference between the classical and our approaches to geodesy is also discussed in this context. The second and third sections describe the milieu of geodetic theory and geodetic practice, as well as the international organizations that cater to geodetic scientists and practicing geodesists. The last section deals with the geodetic profession—its strata and the education of the individual levels of geodesists and surveyors. It also discusses the relation between geodesy and surveying and, in particular, the role geodesy plays in the education of a surveying engineer and surveying technician.

### 4.1. Functions of geodesy

Until a decade or two ago, geodesy was thought to occupy the space delimited by the following definition [HELMERT, 1880, p. 3]: “Geodesy is the science of measuring and portraying the earth’s surface.” Then people involved with geodesy began to realize that this definition no longer fully reflected the role contemporary geodesy played and started searching for a new framework. This search probably culminated in the new definition of geodesy, accepted by the National Research Council of Canada (NRC), that we quote here [ASSOCIATE COMMITTEE ON GEODESY AND GEOPHYSICS, 1973]:

Geodesy is the discipline that deals with the measurement and representation of the earth, including its gravity field, in a three-dimensional time varying space.

At the 1975 Grenoble meeting of the Commission on Education of the International Association of Geodesy (see §4.2), a virtually identical definition [RINNER, 1979] was adopted, except for the inclusion of other celestial bodies and their respective gravity fields.

As with most scientific disciplines, geodesy is arranged into subdisciplines. The classical subdisciplines are: *geometrical geodesy*, *physical geodesy*, *mathematical geodesy*, *dynamic geodesy*. Over the past 30 years or so, new technology and new

applications have given rise to several more 'geodesies'; for example, satellite geodesy, inertial geodesy, marine geodesy, space geodesy, and even horizontal and vertical geodesies. Rather too many geodesies to live with! Although some of these terms appear reasonable, others do not: Are we prepared to call control surveying with a theodolite 'theodolite geodesy'? If we accept vertical geodesy, why not 'oblique geodesy'? Will 'mountain geodesy' take care of the mountains, while 'lowland geodesy' informs us about the remainder of the earth's dry land? Little wonder so many people are bewildered, if not confounded, by geodesy. We believe that the syndrome of too many geodesies is, in part, responsible for the lack of appreciation for the discipline itself. Moreover, one cannot help feeling that it is the geodesists themselves who are mostly responsible for this sorry state of affairs by promulgating this terminology. Be that as it may, the fact still remains that in some parts of the world and by some people, geodesy is being mysticized, while in other parts and by other people, it is being thought of as irrelevant. Neither of these extreme positions is healthy.

We are convinced that the remedy for this situation lies in functionalizing geodesy. This can be accomplished quite naturally when the definition of geodesy is looked into more closely. The result [VANIČEK AND KRAKIWSKY, 1978] is a breakdown into three main functions and, corresponding to them, the following three subdisciplines:

- (a) positioning,
- (b) the earth's gravity field, and
- (c) temporal variations (in positions as well as of the gravity field).

Clearly, in such a functional division of geodesy, there is no room for any artificially defined 'geodesies'; this is the way geodesy is presented in this book.

*Positioning*, or point position determination, is the geodetic task that, for reasons explained in §2.1, the community at large best understands. Points can be positioned either individually or as a part of a whole network of points; the positions sought may be either absolute (with respect to a coordinate system) or relative (with respect to other points). Concepts pertaining to positioning are explained in Part IV of this book.

The reasons why geodesists study the geometry of the *earth's gravity field* were hinted at in §2.2; let us now explain them more fully. The knowledge of the geometry of the gravity field is needed to make possible the transformation of the geodetic observations made in the physical space (affected by gravity) into the geometrical space in which positions are usually defined. In addition, the shapes of equipotential surfaces and plumb lines are needed for projects involving the physical environment (e.g., flow of water). Methods for studying and determining the geometry of the earth's gravity field are shown in Part V of this book.

*Temporal variations of positions and the gravity field* result from deformations of the earth (and its gravity field) attributed to a number of causes. In geodesy, it is immaterial what causes these movements—be it earth tides, crustal loading and rebound, tectonic forces, or other, as yet unknown, phenomena. The study of these causes rightfully belongs to geophysics, but the geometrical aspects fall within the realm of geodesy. This subdiscipline is treated in Part VI of this book.

Others have also functionalized geodesy along lines similar to those described above. For example, the (U.S.) COMMITTEE ON GEODESY [1978; p. 7] states that the major goals of geodesy can be summarized as:

1. Establishment and maintenance of national and global three-dimensional geodetic control networks on land, recognizing the time-variant aspects of these networks.
- 2 Measurement and representation of geodynamic phenomena (polar motion, earth tides, and crustal motion).
3. Determination of the gravity field of the earth including temporal variations.

## 4.2. Geodetic theory

To fulfil all its functions, geodesy must span a spectrum of activities ranging from purely theoretical aspects, needed in laying theoretical foundations for the various geodetic techniques, to field data collection. Accordingly, there are geodesists who specialize in theory and those who specialize in the practice of geodesy. The latter includes fields like control surveys and gravimetry. Of course, the demarcation lines are very blurred and hence defy any firm classifications; nevertheless, some generalizations are possible.

The global nature of geodesy dictates that most of the theoretical work be done either at universities or within governmental institutions. Few private institutes find it economically feasible to do any amount of geodetic research. It is quite usual to combine geodetic theory with practice within one establishment, although specialized geodetic research institutes do exist. Much of geodetic research is also done in the guise of space science, geophysics, oceanography, etc.

Of great importance for geodetic theory is international scientific communication. The communication channels have been secured and formalized under the umbrella of UNESCO's International Council of Scientific Unions. The international organizations in charge of these channels are shown in FIG. 1, according to THE WORLD OF LEARNING 1981-82 [1981]. The organization that is directly responsible for looking after geodetic needs is the *International Association of Geodesy* (IAG) [INTERNATIONAL UNION OF GEODESY AND GEOPHYSICS (IUGG), 1978]. Other international organizations also have a limited vested interest in geodesy as this is more along the engineering and technological lines, these organizations will be introduced in the next section.

The IAG meets every four years, usually together with the other six IUGG associations, to discuss, in the form of scientific symposia, various issues and pass resolutions regarded as recommendations by the member countries. The Association is divided into several commissions, study groups, bureaus, and centres which are set up to deal with contemporary problems and, as such, change from time to time.

Each of the 61 (as of 1981) member countries appoints one official delegate to the IAG. This delegate is usually nominated by the national learned society serving as the professional home for geodesists, and is appointed by the National Academy of

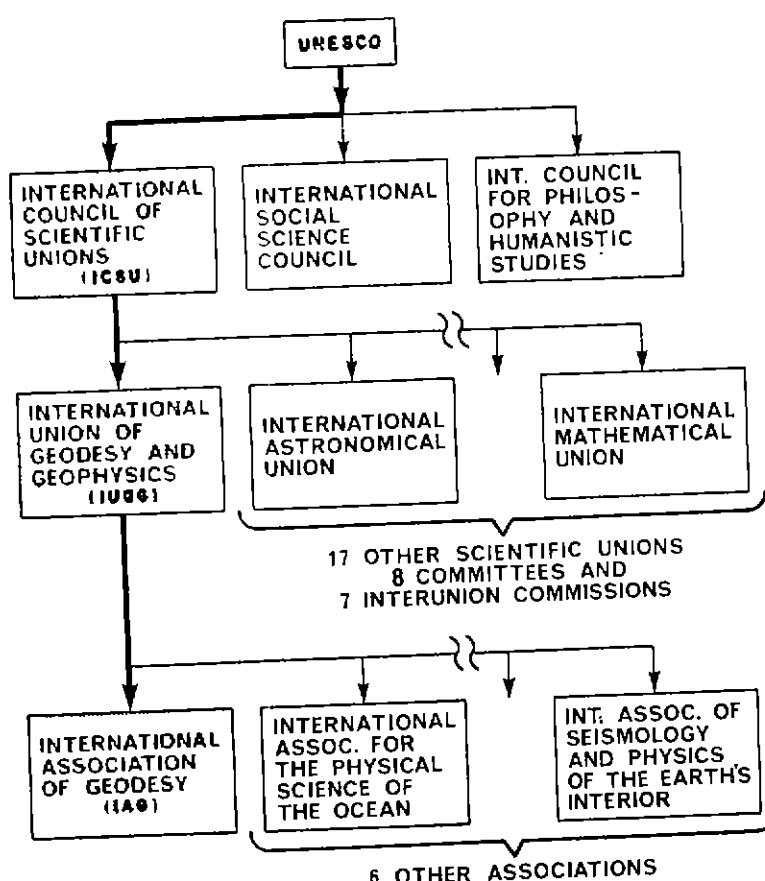


FIG. 4.1. International organization of geodesy.

Science or a similar national institution. Every delegate has one vote in the council of the IAG. To keep each other informed about their geodetic achievements and activities, the member countries submit quadrennial reports to the IAG on the occasion of the IUGG meetings.

The IAG publishes a quarterly scientific journal, *Bulletin Géodésique*, and participates in the publication of an IUGG bimonthly journal of a more administrative nature—the *IUGG Chronical*. In addition to these two official journals, there are many national and international journals catering either fully or partially to geodetic matters.

#### 4.3. Geodetic practice

For reasons explained earlier, the practice of geodesy is quite frequently subjugated to the needs of mapping in individual countries. More often than not, this relation is reflected in the organizational structure of geodesy with the invariable

result that other components of geodetic work are done under the auspices of other professional institutions. For similar reasons, the practice of geodesy in some countries is almost entirely in the hands of the military. While in many cases this proves to be a distinct advantage, in other cases it is to the detriment of the profession, particularly when all the geodetic works are done only in support of military mapping.

Geodetic practice, by its nature, requires not only geodetic professionals—scientists and engineers—but also technicians and auxiliary personnel. It has been estimated [BRANDENBERGER, 1976, 1977] that world geodetic works connected with mapping alone employ some 15 000 professionals, 45 000 technicians, and 90 000 auxiliary personnel. Of these, about 60% are governmental employees, and 40% are from the private sector. The cost of these works totals about \$525 million (in 1976 U.S.\$). Unfortunately, no such estimates exist for other geodetic activities, but it would not be too surprising to find that they consume at least the same amount of money annually.

To achieve its aims, geodesy uses a variety of measuring techniques and systems. They range from simple to complicated, from terrestrial to extraterrestrial, and from purely geodetic to those that are usually recognized as belonging to geophysics, oceanography, or astronomy. The concepts of these techniques and systems will be discussed here, but the description and details are considered outside the scope of this book.

The geodesist practising positioning has, in addition to the IAG, a certain, though limited, international forum in the individual commissions of the *International Federation of Surveyors* (known better under its French name of Fédération Internationale des Géomètres—the FIG); some of these commissions have an interest in geodetic practice. The relation of the FIG to the other international organizations, according to THE WORLD OF LEARNING 1981–82 [1981], is shown in FIG. 2. The functions of the FIG resemble those of the IAG, with a national delegate from each of the 50 nations (in 1981) being nominated by individual national professional

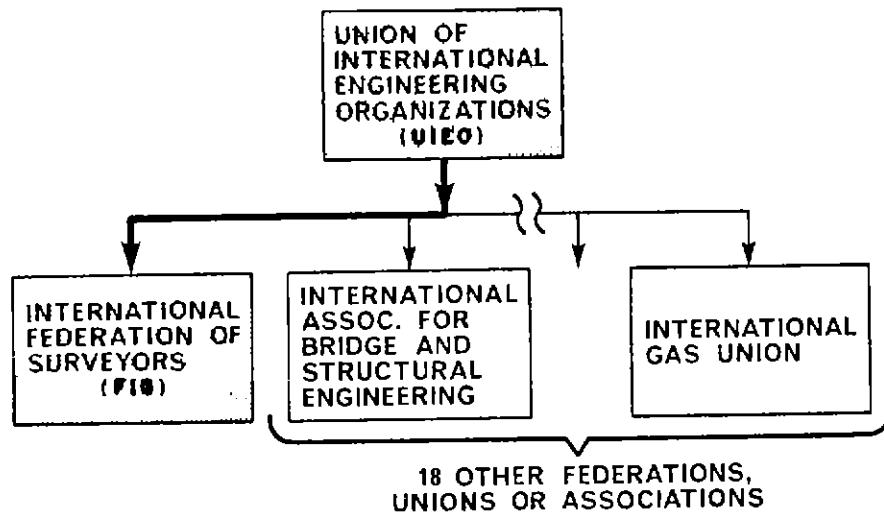


FIG. 4.2. International organization of surveying.

surveying societies directly for each of the separate commissions. The FIG meets every three years and has no official publication of its own. Other international organizations, such as the International Society for Photogrammetry (ISP), the International Cartographic Association (ICA), or the Cartographic Section of the Centre of National Resources, Energy and Transportation of the U.N., also have a distant but distinct interest in geodetic matters.

#### 4.4. Geodetic profession

As stated in the previous section, geodetic personnel may be categorized into scientists, engineers, technicians, and auxiliary personnel. These categories differ by education or experience or both of them. A *geodetic scientist* should typically have a postgraduate degree (a Master or a Doctorate) from a university offering a specialization in geodesy. The extent of the scientist's knowledge should cover at least the topics dealt with in this book and to the same depth.

An engineer is the professional who bridges the gap between the theoretician (scientist) on the one hand and the technician on the other. This person must thus understand the languages of both these groups and be able to communicate freely with them. Specifically, a geodetic or *surveying engineer* should possess an undergraduate degree with geodesy as a major field. The engineer should have a good appreciation of the theory while at the same time have some of the basic skills requisite for the technicians. The surveying engineer should be capable of designing and supervising data collections, carrying out routine data analyses, and even solving smaller problems of a theoretical nature.

Let us pause and have one more look at the relation between geodesy and surveying already touched on in §2.1. Geodesy being the theoretical foundation of surveying means, in practical terms, that a surveying engineer needs to know geodesy in much the same way as an electrical engineer needs to know electricity, a chemical engineer needs to know chemistry, or a mechanical engineer needs to know mechanics. The educational extent to which a surveying engineer should be exposed to geodesy, among other things, should, once more, coincide with the scope of this book. A good understanding of the basics should be required but, compared with a geodetic scientist, the depth in the other topics should be reduced. The lack of the geodetic component in the educational formation reduces a surveying engineer to a surveying technician.

A *surveying technician* (technologist) ideally should have a surveying diploma from a college or technological school. The technician should be well versed in the routine of various kinds of data collection with some understanding of what can be done, or is being done, with the collected data. Thus, among the variety of subjects a technician is taught, only a superficial understanding of geodesy is needed. Nevertheless, there was agreement at the IXth NATIONAL SURVEY TEACHERS' CONFERENCE [1977] that it is more important to give the future technician-technologist an idea of the full spectrum of geodesy than to teach a few tricks and formulae, no matter how well they are selected from the body of geodetic knowledge. Thus Part II of this

book would represent about the right breadth and depth of geodesy for a technological surveying course. In our opinion, the prevalent concept of geodetic education should be that, while the depth of understanding should vary from the scientist to the technician, the breadth should not.

National learned and professional societies may or may not cater to all three groups mentioned above. They usually do not; their interests are deemed sufficiently different to rule this out. The societies are hence normally designed to run along educational rather than professional lines.

Finally, let us state that the employment opportunities for a geodesist are fairly diversified and relatively good at present. We would expect these opportunities to grow in the next few decades. While the geodetic scientists find careers almost exclusively in the academic world and with governmental institutions, surveying engineers and technician-technologists also find equally challenging positions in private enterprise.

## PART I

### REFERENCES

- ABRAMOWITZ, M. AND I.A. STEGUN (Eds.) (1964). *Handbook of Mathematical Functions*. Dover reprint, 1965.
- ASIMOV, I. (1972). *Biographical Encyclopaedia of Science and Technology*. 2nd ed., Avon Books.
- ASSOCIATE COMMITTEE ON GEODESY AND GEOPHYSICS (1973). Minutes of the 60th meeting. National Research Council of Canada, Ottawa.
- BLACHUT, T.J., A. CHRZANOWSKI AND J.J. SAASTAMOINEN (1979). *Urban Surveying and Mapping*. Springer.
- BÖHM, J. (1972). *Vyšší Geodesie I*. ČVUT, Prague, Czechoslovakia.
- BOORSTIN, D.J. (1958). *The Americans: The Colonial Experience*. Random House.
- BOTTING, D. (1973). *Humboldt and the Cosmos*. Sphere Books.
- BRANDENBERGER, A.J. (1976). Study on the status of world cartography. United Nations Economic and Social Council, First United Nations Regional Cartographic Conference for the Americas, Panama.
- BRANDENBERGER, A.J. (1977). Educational trends in the mapping sciences. *Proc. International Symposium on the Changing World of Geodetic Science*, Ed. U.A. Uotila, Columbus, Ohio, October, 1976. Department of Geodetic Science Report 250, Vol. I. The Ohio State University, Columbus, U.S.A., pp. 22–33.
- BROWN, L.A. (1949). *The Story of Maps*. Bonanza Books.
- BUNBURY, E.H. (1883). *A History of Ancient Geography*. Vol. 1, Dover reprint, 1959.
- CHURCHILL, R.V. AND J.W. BROWN (1974). *Complex Variables and Applications*. 3rd ed., McGraw-Hill.
- CLARK, R.W. (1971). *Einstein, the Life and Times*. Nelson, Foster and Scott.
- COMMITTEE ON GEODESY (1978). Geodesy: Trends and prospects. U.S. National Research Council, Washington, D.C., U.S.A.
- COOK, A.H., D.G. KING-HELE, S.A. RAMSDEN AND A.R. ROBBINS (organizers) (1977). A discussion on methods and applications of ranging to artificial satellites and the moon. *Philos. Trans. Roy. Soc. London Ser. A* 284 (1326), pp. 419–619.
- DAVIES, P.C.W. (1979). Einstein's legacy. *The Sciences* 19 (3), pp. 25–28.
- DUKSTERHUIS, E.J. (1950). *The Mechanization of the World Picture*. Translation by C. Dikshoorn of 1950 Dutch edition, Oxford University Press, 1961.
- DREYER, J.L.E. (1905). *A History of Astronomy from Thales to Kepler*. 2nd ed., Dover reprint, 1953.
- DURANT, W. (1944). *The Story of Civilization*. Simon and Schuster, 11 vols.
- Encyclopaedia Britannica (1970). VOL. 10.
- FADDEEV, D.K. AND V.N. FADDEEVA (1963). *Computational Methods of Linear Algebra*. Translated from Russian by R.C. Williams, Freeman.
- FITE, E.D. AND A. FREEMAN (1926). *A Book of Old Maps Delineating American History*. Dover reprint, 1969.
- FLANDERS, H. AND J.J. PRICE (1978). *Calculus with Analytic Geometry*. Academic Press.
- FREUND, J.E. (1971). *Mathematical Statistics*. 2nd ed., Prentice-Hall.
- GREENBERG, M.D. (1971). *Application of Green's Functions in Science and Engineering*. Prentice-Hall.
- GROUEFF, S. (1974). *L'Homme et la Terre*. Larousse.
- HAGIHARA, Y. (1971). *Perturbation Theory*. Vol. II of *Celestial Mechanics*, The MIT Press.

- HAMILTON, A.C. (1969). Problems in land registration and in filing environmental data in eastern Canada. *Canad. Surv.* 23 (1), pp. 12-29.
- HANCOCK, H. (1917). *Theory of Maxima and Minima*. Dover reprint, 1960.
- HAPGOOD, C.H. (1966). *Maps of the Ancient Sea Kings*. Chilton.
- HELMERT, F.R. (1880). *Die mathematischen und physikalischen Theorien der höheren Geodäsie*. Vol. I, Minerva G.M.B.H. reprint, 1962.
- HIEBER, S. AND T.D. GUYENNE (Eds.) (1978). *Proceedings of the European Workshop on Space Oceanography, Navigation and Geodynamics*. ESA, Council of Europe, EARSeL, Schloss Elmau, Germany, January. European Space Agency Report ESA SP-137, Paris, France.
- HOGG, R.V. AND A.T. CRAIG (1970). *Introduction to Mathematical Statistics*. 3rd ed., Macmillan.
- HOHEISEL, G. AND A.M. TROPPER (1963). *Integral Equations*. Translated from German by W. de Gruyter and Co., Berlin, 1968.
- INTERNATIONAL UNION OF GEODESY AND GEOPHYSICS (1978). IUGG year book 1978. *IUGG Chronicle* 126/127, May.
- JACCHIA, L.G. AND J.W. SLOWEY (1975) A catalogue of atmospheric densities from the drag on five balloon satellites. Smithsonian Astrophysical Observatory Special Report 368, Cambridge, U.S.A.
- JASWON, M.A. AND G.T. SYMM (1977). *Integral Equation Methods in Potential Theory and Elastostatics*. Academic Press.
- KOCHIN, N.E. (1961). *Vektornoe Ischislenie I Nachala Tenzornogo Ischislenija*. 8th ed., Publishing House of the USSR Academy of Sciences.
- KORN, G.A. AND T.M. KORN (1968). *Mathematical Handbook for Scientists and Engineers*. 2nd ed., McGraw-Hill.
- KRAKIWSKY, E.J. AND P. VANÍČEK (1974). Geodetic research needed for the redefinition of the size and shape of Canada. *Proc. Geodesy for Canada Conference*, National Advisory Committee on Control Surveys and Mapping, Ottawa, Canada, January. Surveys and Mapping Branch of the Department of Energy, Mines and Resources, Ottawa, Canada.
- KREYSZIG, E. (1978). *Introductory Functional Analysis with Applications*. Wiley.
- LANDKOF, N.S. (1972). *Foundations of Modern Potential Theory*. Springer.
- LENNON, G.W. (1974). Mean sea level as a reference for geodetic levelling. *Canad. Surv.* 28 (5), pp. 524-530.
- MACMILLAN, W.D. (1930). *The Theory of the Potential*. Dover reprint, 1958.
- MCCONNELL, A.J. (1931). *Applications of Tensor Analysis*. Dover reprint, 1957.
- MORRISON, N. (1969). *Introduction to Sequential Smoothing and Prediction*. McGraw-Hill.
- MUELLER, I.I. (Ed.) (1978). Applications of geodesy to geodynamics. *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference*, IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280, The Ohio State University, Columbus, U.S.A.
- NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (1978). NASA directory of station locations. 4th ed., prepared by Computer Sciences Corporation for Goddard Space Flight Center, Greenbelt, U.S.A.
- NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (1979). Application of space technology to crustal dynamics and earthquake research. NASA Technical Paper 1464, Washington, D.C., U.S.A.
- IXTH NATIONAL SURVEYING TEACHERS' CONFERENCE (1977). *Proceedings*, UNB, Fredericton, N.B., June. Department of Surveying Engineering, University of New Brunswick, Fredericton, Canada.
- NORDENSKJÖLD, A.E. (1889). *Facsimile-Atlas to the Early History of Cartography*. Dover reprint, 1973.
- PANNEKOEK, A. (1951). *A History of Astronomy*. Translation of 1951 Dutch ed., George Allen and Unwin, 1961.
- PROTTER, M.H. AND C.B. MORREY JR. (1973). *Calculus for College Students*. 2nd ed., Addison-Wesley.
- RFTORYS, K. (Ed.) (1969). *Survey of Applicable Mathematics*. Translated from Czech by Dr. Rudolf Vyborný et al., 1968, The MIT Press.
- RINNER, K. (1979). Report of the IAG Commission IX (education). Paper presented at the XVII IUGG General Assembly, Canberra, Australia.
- SOUTH, E.J. (1884). *Dynamics of a System of Rigid Bodies*. Part II, 4th ed., Dover reprint, 1955.
- SAVAGE, J.C. AND R.O. BURFORD (1973). Geodetic determination of relative plate motion in central California. *J. Geophys. Res.* 78, pp. 832-845.

## REFERENCES, PART I

- SYNGE, J.L. AND A. SCHILD (1949). *Tensor Calculus*. University of Toronto Press.
- TELFORD, W.M., L.P. GELDART, R.E. SHERIFF AND D.A. KEYS (1976). *Applied Geophysics*. Cambridge University Press.
- THOMPSON, E.H. (1969). *An Introduction to the Algebra of Matrices with some Applications*. University of Toronto Press.
- THOMSON, D.W. (1966). *Men and Meridians*. Vol. 1, Queen's Printer, Ottawa.
- TOMPKINS, P. (1971). *Secrets of the Great Pyramid*. Appendix by L.C. Stecchini. Harper and Row.
- VANÍČEK, P. (1976). Papel de la geodesia en la sociedad. *Proc. Annual Meeting of the National Congress of Photogrammetry, Photointerpretation and Geodesy*, Mexico City, Mexico, May, pp. III-1-III-9.
- VANÍČEK, P. (1977). Geophysical applications of geodesy. *Proc. Symposium of the Geophysics Commission of the Pan American Institute of Geography and History*, Ed. J.G. Tanner and M.R. Dence. Ottawa, Canada, September, 1976. Publication of the Earth Physics Branch of the Department of Energy, Mines and Resources, Ottawa, Vol. 46, No. 3, pp. 45-48.
- VANÍČEK, P. AND E.J. KRAKIWSKY (1978). Geodesy reborn. *Surveying and Mapping XXXVII* (1), pp. 23-26.
- WELLS, H.G. (1961). *The Outline of History*. Vols. I, IV, Garden City Books.
- WILLIAMSON, R.E., R.H. CROWELL AND H.F. TROTTER (1972). *Calculus of Vector Functions*. 2nd ed., Prentice-Hall.
- WONNACOTT, T.H. AND R.J. WONNACOTT (1972). *Introductory Statistics*. 2nd ed., Wiley.
- World of Learning 1981-82, The (1981). Vol. 1. 29th ed., Europa Publications.
- WREDE, R.C. (1963). *Introduction to Vector and Tensor Analysis*. Dover reprint, 1972.
- WYLIE, C.R., JR. (1966). *Advanced Engineering Mathematics*. 3rd ed., McGraw-Hill.

**PART II**

**THE EARTH**

## CHAPTER 5

### EARTH AND ITS MOTIONS

Since various kinds of extraterrestrial measurements—such as astronomical (both optical and radio), satellite, and lunar—play a major role in geodesy, it is necessary to develop some understanding of how the earth moves among other celestial bodies. It is known that the earth undergoes the following kinds of motions simultaneously:

- (a) It moves with our galaxy in respect to other galaxies.
- (b) It circulates with the solar system within our galaxy.
- (c) It revolves around the sun, together with other planets.
- (d) It rotates (spins) around its instantaneous axis of rotation.

Of these motions, the first two are of importance to astronomers studying galactic and intergalactic phenomena. When dealing only with the earth, we generally do not have to worry about them because most of the celestial objects used for our observations are well within our galaxy. Thus our concentration will be on the last two motions—the annual (around the sun), and the diurnal (around the earth's axis of spin).

Interestingly, two radically different physical concepts are used to describe the annual and diurnal motions. The annual motion can be adequately explained using celestial mechanics; i.e., by regarding the earth and other celestial bodies as 'point masses', or particles without dimensions. To explain the diurnal motion with its side effects of precession and nutation, the earth has to be regarded as a massive body or as a gyroscope. These two dynamic models of the earth are treated in the first two sections of this chapter.

In the last two sections, polar motion, a complication that stems from the earth's diurnal spin, is introduced. This particular motion has profound implications in geodesy. To adequately explain the motion, the model of the earth has to be further refined by taking into account its rheology (behaviour under stress), atmosphere, oceans, etc.

#### 5.1. Earth's annual motion

In a description of the annual motion, the dimensions of the earth and other celestial bodies can be regarded as negligible, compared with the dimensions of the solar system. Under these conditions, Kepler's three laws apply as follows

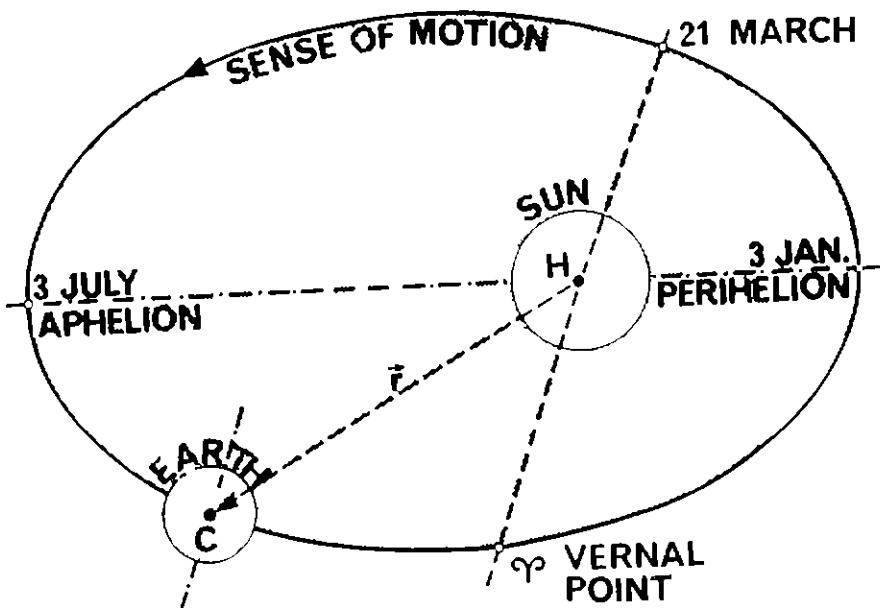


FIG. 5.1. Annual motion of the earth.

[KOVALEVSKY, 1967]:

- (a) The orbit of any planet is an ellipse and the sun (*H*) stands in one of its foci.
- (b) A planet moves along its orbit with a constant areal velocity; i.e., the area swept by the radius vector  $\vec{r}$  of the planet (see FIG. 1) is constant for a given time interval.
- (c) The ratios of squares of orbital periods (*m*) of the planets are the same as the ratios of cubes of lengths of major semi-axes ( $a_0$ ) of their orbits:

$$m^2/a_0^3 = \text{const.} \quad (5.1)$$

The plane of the earth's orbit is called the *ecliptic*. Because of Kepler's second law, the earth moves faster when it is closer to the sun and slower when it is farther away. It completes one revolution (with respect to the stars) in one *sidereal year*. In reality, the presence of other planets and the moon influence (perturb) the shape of the earth's orbit, so it is not exactly elliptical and not even planar. These perturbations are, however, very small compared with the orbit's dimensions and, for many practical purposes, can be neglected.

The point of closest approach to the sun (see FIG. 1) is called the *perihelion*, and is located at one end of the major orbital axis. The other end point is called the *aphelion*, which is obviously the point of farthest recession. Another important point on the orbit is the vernal point ( $\gamma$ ), which will be defined in the next section: an understanding of the diurnal motion is necessary before it can be defined.

At present, the earth passes through the perihelion around January 3 and through the aphelion around July 3 [NASSAU, 1948]. These change by a few days every year. The whole orbital ellipse moves with respect to the surrounding stars in the galaxy, but the movement is so slow it can be disregarded for most tasks.

### 5.2. Earth's spin, precession, and nutation

In describing the earth's spin, the earth's dimensions can no longer be neglected. In the next dynamically simplest model, the earth is taken as a rigid body which, while travelling around the sun, spins around an axis passing through the body. In mechanics, such a body is referred to as a *gyroscope*.

As known from everyday life, the main manifestations of the earth's gyroscopic motion—the diurnal rotation around the earth's polar axis—is the occurrence of the days and nights. It takes the earth 366.2564 rotations with respect to the stars—known as *sidereal days*—or 365.2564 rotations with respect to the sun—known as (mean) *solar days*—to complete one revolution around the sun (one sidereal year). To a high degree of accuracy, the spin axis, called the *instantaneous spin axis*, coincides with the earth's principal axis of the maximum moment of inertia that passes through the earth's centre of mass (*C*). As will be seen in the following two sections, the difference between these two axes is significant from the geodetic point of view and thus will have to be treated in detail.

When there is an external torque exerted on the spinning gyroscope, the spin axis of the gyroscope describes a circular cone with its vertex located at the centre of mass of the gyroscope. This motion is known as *precession*. In the case of the earth, it is the attraction of celestial bodies that supplies the torque; the situation for the sun is depicted in FIG. 2. It is clear that the hemisphere closer to the sun is attracted more than the one farther away. To obtain the torque as shown, i.e., the torque with respect to *C*, the reference point for describing precession, we must first subtract the force acting at *C* from both of these hemispherical forces. (The same situation exists

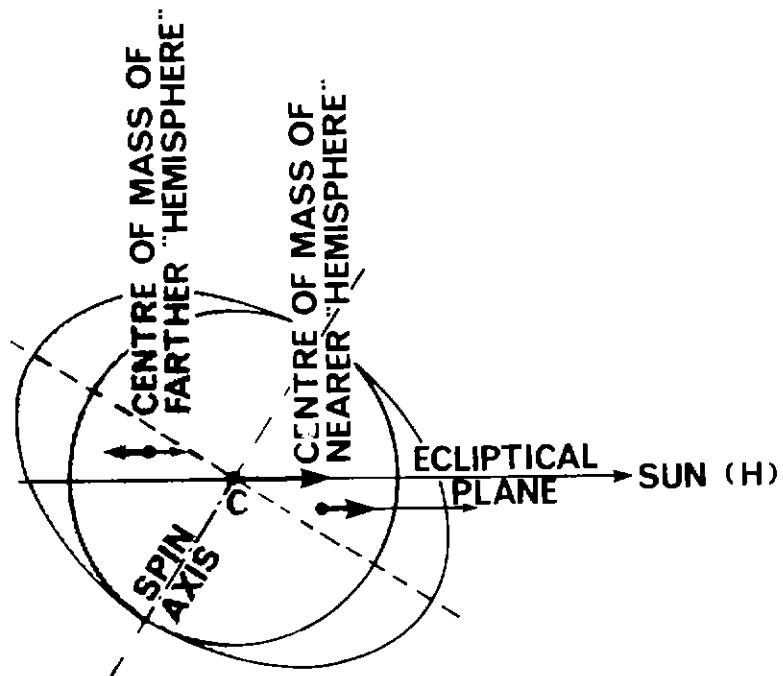


FIG. 5.2. Solar precession torque.

with tidal forces, as we will see in §8.1.) The presence of the (equatorial) bulges is necessary for the precession to arise; if they were absent, the torque would disappear because the forces in the couple would both lie on the *C-H* line.

If the centre of mass of the gyroscope is made to move in a plane, then the precession axis (i.e., the axis of the precession cone) is perpendicular to this plane. This is the case of the earth, with the sun and moon exerting the bulk of the torque. The orientation of the earth's spin axis is not fixed in space but moves slowly along a cone perpendicular to the ecliptical plane (FIG. 3). It completes one precession cycle in approximately 26 000 years; this period is known as the Platonic year.

The spin axis is inclined with respect to the ecliptic by an almost constant angle of about  $66.5^\circ$ , so the angle  $\epsilon$  between the equatorial and ecliptical planes is about  $23.5^\circ$ . Evidently, if it were not for this inclination, called *obliquity*, the earth would spin around an axis perpendicular to the ecliptical plane, and the days and nights would be equally long—12 hours—at any time and any place on the earth. Because of the obliquity, the sun shines longer on the Northern Hemisphere for one-half of the earth's revolution, and then on the Southern Hemisphere for the second half. Logically then, there are two points on the earth's orbit where the sun shines on both hemispheres for the same length of time (i.e., 12 (solar) hours). The dates corresponding to these two points are thus known as equinoxes, meaning 'equal nights'.

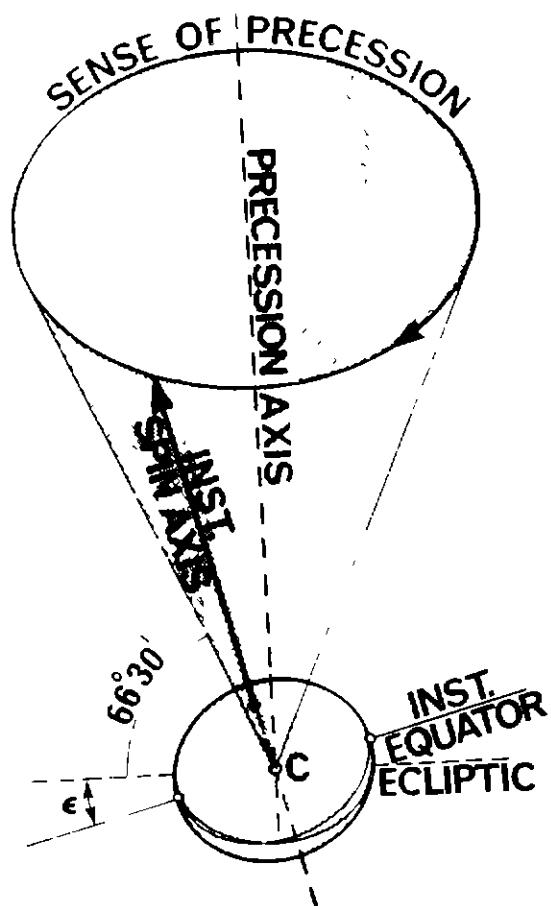


FIG. 5.3. Precession cone.

The dates corresponding to the two equinoxes are determined simply by observing the times of the sunrise and sunset. This is the reason why one of them, the vernal equinox (when spring comes to the Northern Hemisphere), has been selected as the reference that serves as the origin for various measurements on the ecliptic. The line connecting the centres of the earth and the sun at the instant of the vernal equinox is the intersection of equatorial and ecliptical planes. It points in an almost constant direction among the stars. Hence to an observer on the earth, the sun at the vernal equinox always appears at a specific point among the stars. This particular point is referred to as the *vernal point*, or vernal equinox (cf. FIG. 1).

It is not difficult to see that the vernal point moves as the precession progresses. It describes one full circle in about 26 000 years, which means that it moves along the ecliptic by approximately  $0.014^\circ$  per year. Precise astronomical observations have shown the rate of motion to be  $50.3''$  per year [MUELLER, 1969], which gives a more accurate length of 25 765 years to the precession period (Platonic year). An uncertainty of  $0.1''$  in the observed rate of motion corresponds to an uncertainty of about 50 years in the length of this period. A mental exercise, using FIGS. 1 and 3, convinces us that the motion of the vernal point along the ecliptic is clockwise, i.e., against the annual motion of the earth.

The presence of the moon makes the study of the earth's kinematics more engaging. The first important fact is that the moon orbits the earth on a plane that is inclined, with respect to the ecliptical plane, by  $5^\circ 11'$  [MUELLER, 1969]. The intersection of the lunar orbital plane with the earth's ecliptical plane, known as the *nodal line* (FIG. 4), rotates once in every 18.6 years. This introduces a periodic change in the external torque. Thus the moon, in addition to perturbing the earth's annual orbit, also perturbs the precession. This additional perturbation results in another motion of the earth's spin axis, called *forced nutation* or simply *nutation*.

The nutation cone is much narrower than that of the precession. Its vertex angle is only  $18.42''$  as compared with  $47^\circ$  for precession. Also, nutation is obviously much

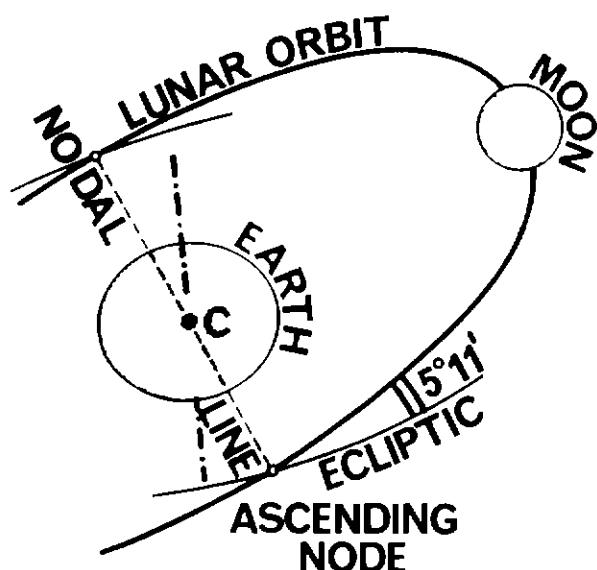


FIG. 5.4. Lunar orbit.

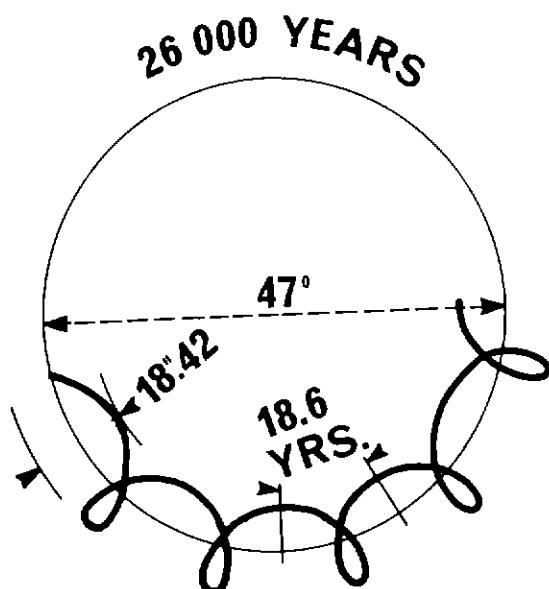


FIG. 5.5. Precession and nutation of the earth's spin axis.

faster than precession, completing one full circle in 18.6 years as compared with 26000 years. The composite movement of the earth's spin axis with respect to the ecliptic is shown diagrammatically in FIG. 5. Obviously, the nutation also influences the position of the vernal point, but the effect is minute.

The mathematical description of the gyroscopic motion, precession and nutation, is tedious. Since the luni-solar torque that causes these motions is a function of the positions of the moon and sun, it changes constantly. In the composite motion, i.e., in the *luni-solar precession* plus nutation, there are well-defined periodicities due to the luni-solar positions. Among these, the two above-mentioned periods (18.6 and 26000 years) are just the principal ones, whose contributions have the largest amplitudes. Other periods that contribute appreciably are (a) semiannual, with amplitude of the order of 0.5" (compared with 9.21" for 18.6 years); and (b) fortnightly, whose amplitude reaches about 0.1" [MELCHIOR, 1973]. Some of these periods exist in the tidal potential, which will be described in §8.1.

The position and orientation of the earth in space, at any time, is the result of all the above described motions. Therefore, all these motions have a direct effect on the astronomical and satellite observations conducted from the earth, and as such must be taken into account. The manner in which these motions have to be accounted for is adequately covered in various textbooks (e.g., in KAULA [1966b] and MUELLER [1969]), yearbooks, and star catalogues listing the astronomical coordinates of stars used for geodetic observations. Similarly, the detailed theory of the motions, discussed so far, is readily available in many textbooks on celestial mechanics (e.g., in NEWCOMB [1906], SMART [1956], and MELCHIOR [1973]). Hence, this subject will not be further explored here. The motion of the instantaneous axis of rotation with respect to the earth, though, is a different matter. In this case, the motion concerns the earth more intimately, hence this motion will be considered, in some detail, in the next two sections.

### 5.3. Earth's free nutation

Dynamically, the motion known as *earth's free nutation*, or wobble, is a torque-free nutation—an effect that accompanies any gyroscopic motion. To derive the differential equations of the wobble, let us first adopt the most suitable natural system of coordinates for this task. By 'natural' is meant a system that is dictated by some physical properties of the earth and that is independent of any subjective preferences. The natural system to use here is the geocentric system, which is related to the principal moments of inertia. Its axes are given by the eigenvectors of the earth's tensor of inertia  $\mathbf{J}$  [SYMON, 1971]. In other words, it is a right-handed ( $x$ ,  $y$ ,  $z$ ) Cartesian system with its origin at the centre of mass  $C$  of the earth and its axes coincident with the axes of the principal ellipsoid of inertia—not to be confused with the ellipsoid that approximates the shape of the earth—so that the  $z$ -axis points north (FIG. 6). For a rigid body, this coordinate system is rigidly linked to the body. For a non-rigid body, the position of the system, at any time, is given by the instantaneous distribution of the mass within the body.

Denoting the three principal moments of inertia with respect to  $x$ ,  $y$ ,  $z$  by  $I_1$ ,  $I_2$ ,  $I_3$  respectively, *Euler's equation* for free nutation reads [MACMILLAN, 1936]

$$\mathbf{J}\dot{\boldsymbol{\omega}} + \bar{\boldsymbol{\omega}} \times \mathbf{J}\bar{\boldsymbol{\omega}} = \bar{0}, \quad (5.2)$$

or

$$\begin{bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{bmatrix} \begin{bmatrix} \dot{\omega}_1 \\ \dot{\omega}_2 \\ \dot{\omega}_3 \end{bmatrix} + \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \begin{bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} = \bar{0},$$

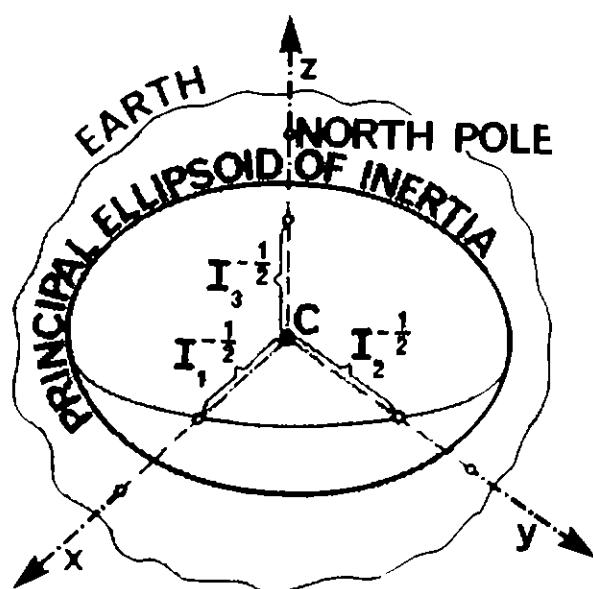


FIG. 5.6. Principal ellipsoid of inertia of the earth.

where  $\omega_1, \omega_2, \omega_3$  are the components of the instantaneous spin angular velocity vector  $\bar{\omega}$  in the  $x, y, z$  system, and  $\dot{\omega}_1, \dot{\omega}_2, \dot{\omega}_3$  are their respective time derivatives. Obviously, the vector  $\bar{\omega}(\tau)$  from (2), taken as a function of time  $\tau$ , defines the position of the instantaneous axis in our geocentric natural system.

Equation (2) is a vector differential equation of first order (see §3.2). Observational evidence shows that the earth's two equatorial moments of inertia  $I_1, I_2$ , are, to a high degree of accuracy, equal to each other but significantly different from the polar moment of inertia  $I_3$ . By putting  $I_1 = I_2$  in (2), the following three ordinary differential equations (see §3.2) are obtained that describe the free nutation:

$$\dot{\omega}_1 + \frac{I_3 - I_1}{I_1} \omega_2 \omega_3 = 0, \quad \dot{\omega}_2 - \frac{I_3 - I_1}{I_1} \omega_1 \omega_3 = 0, \quad \dot{\omega}_3 \doteq 0. \quad (5.3)$$

From the third equation, it is seen that the polar component

$$\omega_3(\tau) \doteq \text{const.} \quad (5.4)$$

Let us denote this constant by  $\mu$ . By differentiating (3) with respect to time, the first two equations can be transformed to the following second-order differential equations,

$$\ddot{\omega}_1 + \left( \frac{I_3 - I_1}{I_1} \right)^2 \mu^2 \omega_1 = 0, \quad \ddot{\omega}_2 + \left( \frac{I_3 - I_1}{I_1} \right)^2 \mu^2 \omega_2 = 0. \quad (5.5)$$

Note that in these equations the variables are separated.

Equations (5) describe simple harmonic motion (cf. §3.2). Taking into account the first-order differential equations (3), the two equatorial components  $\omega_1, \omega_2$  can be written as

$$\begin{aligned} \omega_1(\tau) &= \beta \cos \left( \frac{I_3 - I_1}{I_1} \mu \tau + \psi \right), \\ \omega_2(\tau) &= \beta \sin \left( \frac{I_3 - I_1}{I_1} \mu \tau + \psi \right), \end{aligned} \quad (5.6)$$

where  $\beta, \psi$  are integration constants with arbitrary values. A brief look at the three components of  $\bar{\omega}(\tau)$  in  $xy$ ,  $xz$ , and  $yz$  planes (FIG. 7), shows that the instantaneous spin axis again describes a circular cone around the earth's polar, principal axis of inertia. It can be shown from (6) that the sense of the motion is anti-clockwise when viewed from the north pole. It is this instantaneous axis that precesses and nutates and whose motion was described in §5.2. Also note that, mathematically, even the precession is described by the Euler equations when the external torque replaces the zero-vector on the right-hand side.

Euler's equations and their solutions cannot enlighten us about the values of any of the integration constants, i.e., the vertex angle  $\beta$ , or the phase angle  $\psi$ , or the

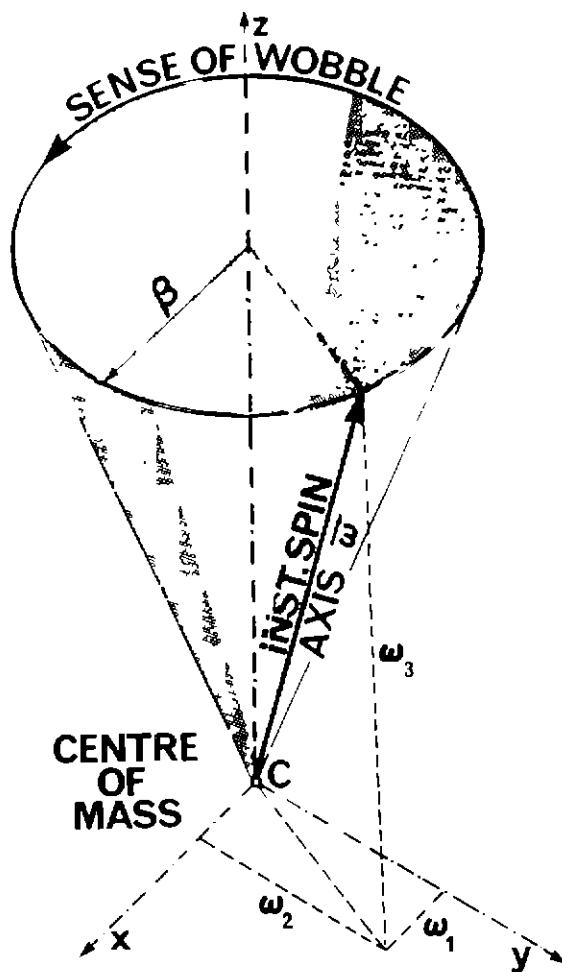


FIG. 5.7. Instantaneous spin axis in geocentric natural coordinate system.

period of the wobble. These have to be determined from observations. Such observations indicate that the  $\omega_1, \omega_2$  components are very small compared with  $\omega_3$ . In other words, the instantaneous axis of spin deviates from the earth's polar axis of inertia by a very small angle ( $\beta$ ), so  $\omega_3$  can be regarded as being practically equal to the magnitude of  $\bar{\omega}$ . This means that  $\omega_3$  can be taken as being practically equal to the *earth's angular velocity* (frequency):

$$\omega_3 = \mu = \omega = 2\pi/1 \text{ sidereal day}. \quad (5.7)$$

Since the free nutation period  $P$  is equal to  $2\pi/\text{frequency}$ , and the frequency is equal to  $((I_3 - I_1)/I_1)\mu$  (cf. eqn. (6)), then

$$P = 2\pi \frac{I_1}{(I_3 - I_1)\mu}. \quad (5.8)$$

Taking the value 305, as determined from precession and nutation observations, for the reciprocal *dynamic flattening*  $H^{-1} = I_1/(I_3 - I_1)$  [MELCHIOR, 1966],  $P$  equals 305 sidereal days. This value is usually referred to as the *Euler period*. (Note that other authors, e.g., HEISKANEN AND MORITZ [1967], define the dynamic flattening as  $(I_3 - I_1)/I_3$ .) When the first accurate, observational data for the wobble became

available at the end of the last century, it was discovered that the actual period was about 40% longer than the Euler period [CHANDLER, 1891]. This discrepancy, as subsequently explained by NEWCOMB [1892], arises because the earth is not rigid. The non-rigidity of the earth tends to increase the wobble period, which is now known to be about 435 solar days and is called the *Chandler period* [ROCHESTER, 1973].

Upon the realization that the earth behaves like a deformable, rather than rigid, body, it became necessary to theoretically account for the internal friction, and thus the dissipation of energy, within the earth. But, whenever the energy of a dynamic system is dissipated, it results in the damping of the motion of the system. Therefore, in our case, we theoretically should expect the amplitude  $\beta$  of the wobble to be damped, i.e., to decrease exponentially with time. The mathematical description of the free nutation of a deformable body is no longer given by Euler's equation but by *Liouville's equation*—see, e.g., MUNK AND MACDONALD [1960].

#### 5.4. Observed polar motion and spin velocity variations

In order to determine the unknown parameters of the polar wobble, the International Astronomical Union (IAU—see FIG. 4.1) set up a programme—International Latitude Service (ILS)—to observe the *actual polar motion*. As will be seen in §15.2, the variations of the pole position are directly observable as the variations in latitude. Five stations (Mizusawa, Japan; Kitab, Russia; Carloforte, Italy; Gaithersburg, and Ukiah, U.S.A.), located nearly on the same parallel of latitude  $39^{\circ}08'$  north, began observing the phenomenon in 1899. Since then, the network of permanent observing stations has grown to over one hundred. They now operate under the auspices of two agencies—the International Polar Motion Service (IPMS) with headquarters in Mizusawa, and the Bureau International de l'Heure (BIH) with headquarters in Paris. In 1969, the United States Naval Weapons Laboratory began to operate their own polar motion service—Dahlgren Polar Monitoring Service (DPMS)—based on their worldwide network of TRANSIT satellite stations (see §15.3).

Thus at present, there is a wealth of observational material consisting of long and comparatively homogeneous observational series. What do these observed series reveal? To begin with, it is now known that the actual polar motion is much more complicated than originally thought. FIG. 8 shows the actually observed motion, during the period 1962 to 1977, with respect to the *Conventional International Origin* (CIO), which is defined as the mean position of the instantaneous pole during the period 1900 to 1905. At first glance the motion does not show any sign of being damped. The accepted explanation is that besides damping there must exist a mechanism that occasionally or continually excites the wobble. The hypothesis that the *wobble excitation* is linked with tectonic earthquakes [MANSINHA AND SMYLIE, 1967] seems, so far, to be the only realistic one. Yet, the quantitative estimates of parameters connected with the excitation and damping have been, until now, very imprecise and not very convincing [JEFFREYS, 1970; PEDERSEN AND ROCHESTER, 1972].

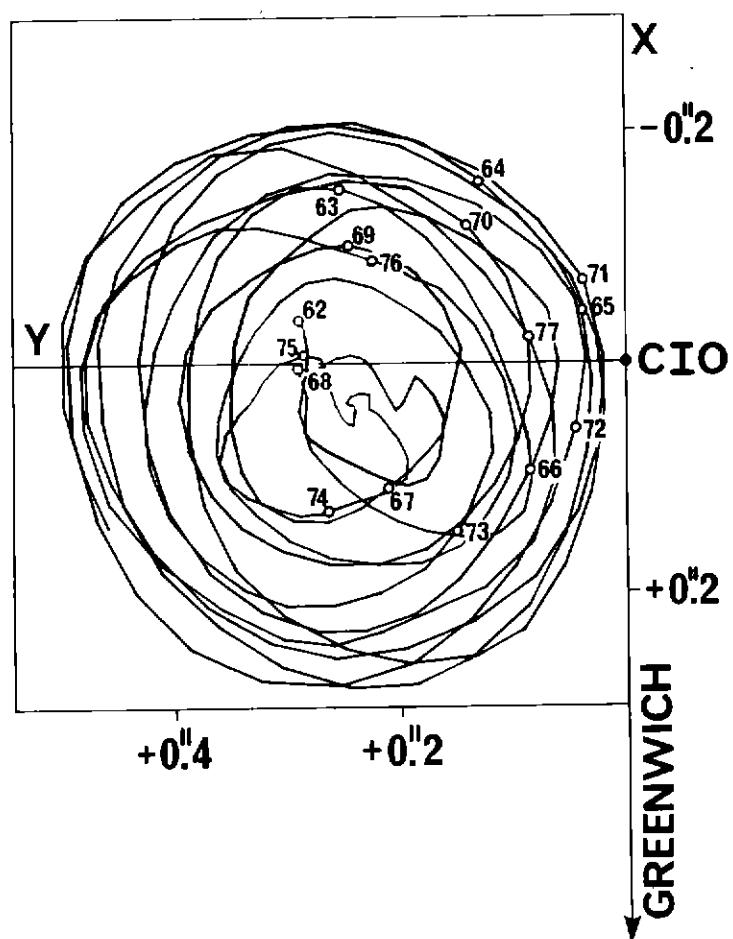


FIG. 5.8. Observed polar motion. (Courtesy of Dr. S. Yumi [1977], Director of IPMS.)

The determination of the most probable value for the Chandler period has been the topic of numerous investigations. Results of some of the more recent attempts are listed in TABLE I.

It is obviously difficult to come up with a firm figure for the amplitude  $\beta$  of the wobble because of the damping and excitation. Indeed, the various analyses of the collected data show a large spread between 0.1" and 0.2". These values correspond

**TABLE 5.1**  
Some results of determination of the Chandler period

Solution	Chandler period (Solar days)	Span of data	Source of data
JEFFREYS [1968]	433.15	1899-1967	ILS
VANIČEK [1969]	435.1	1951-1966	BIH
YUMI [1970]	429.9	1890-1969	ILS
	439.4	1963-1969	ILS
ANDERLE [1970]	416.6	1967-1970	DPMS
CURRIE [1974]	432.95	1900-1973	ILS
GRABER [1976]	430.8	1960-1974	IPMS

to an actual displacement of the instantaneous pole, on the surface of the earth, of 3 to 6 metres.

Various analyses performed on the existing data have also revealed the presence of two more significant components of the actual polar motion—seasonal and secular variations. The exact mechanism controlling these two components is, as yet, not very well understood. The seasonal variation (with an annual period), superimposed on the Chandler path, displays considerable fluctuation in amplitude. According to ORLOV [1961], the amplitude ranges from  $0.04''$  to  $0.12''$ , which corresponds to a pole displacement of 1 to 4 metres. The variation is probably intimately related to the seasonal variation in temperature, barometric pressure, snow load, precipitation, etc. [MUNK AND MACDONALD, 1960]. Others (e.g., VANÍČEK [1971], WELLS AND CHINNERY [1973]) believe that a part of the seasonal variation is only spurious, appearing in the observed polar path as a result of different seasonal effects experienced by individual observing stations.

Perhaps even less understood is the secular variation. It manifests itself as a drift of the pole of about  $0.002''$  to  $0.003''$  per year. The explanation for this drift is thought to be in tectonic movements (see §8.3). Also, some investigators claim to have detected the presence of a long-period wiggle of about 24 years [MARKOWITZ AND GUINOT, 1968; VANÍČEK, 1969]. In any case, it is difficult to see to what extent these superimposed variations are related to the mechanics of free nutation.

In §5.2 it was stated that the earth makes approximately 366.2564 rotations (sidereal days) while completing one revolution around the sun (sidereal year). The velocity of the earth's rotation had been considered constant in the past, so constant in fact that until the 1930s it was universally accepted as the best time-keeping device. However, an increase in the accuracy of both the observations and the clocks has revealed that variations of the earth's spin velocity do exist. With the introduction of the BIH, and the advent of atomic clocks in 1955, our appreciation of these variations has further increased.

At present, three kinds of *spin velocity fluctuation* are recognized [MARKOWITZ, 1972]: secular, periodic, and irregular. The continuous (secular) slowing down of the earth's spin, mainly due to tidal friction, leads to an increase in the length of the day (l.o.d.) by 2 milliseconds per century. Seasonal (with periods of one year and half a

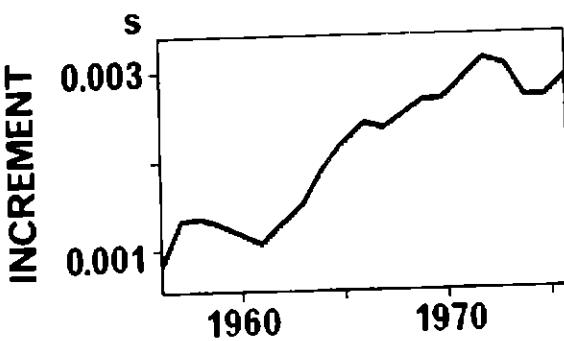


FIG. 5.9. Variation in the earth's spin velocity in terms of the length of the day (86400 s plus increment) with periodic terms removed. (Courtesy of Dr. B. Guinot [1977], Director of BIH.)

year), and other periodic variations (with periods of around one month), may attain magnitudes of the order of several milliseconds. They are due partly to tidal and partly to wind effects [MUNK AND MACDONALD, 1960]. A high degree of correlation exists between the I.O.D. and the atmospheric angular momentum at a period near 50 days [ANGLEY ET AL., 1981]. The most interesting observed phenomena relating to the earth's spin are the irregular, abrupt changes in the rotation rate. They can amount to 10 milliseconds per day. If real, these changes would indicate some, as yet unknown, excitation mechanism. Contrary to the case of polar wobble, earthquakes do not seem to be a possible explanation [MARKOWITZ, 1972]. The actually observed fluctuations, during the period 1959 to 1976, are shown in FIG. 9.

## CHAPTER 6

### EARTH AND ITS GRAVITY FIELD

Instruments with which geodetic measurements are made, on and above the surface of the earth, are subjected to various physical forces. To interpret the results of the measurements properly, it is necessary to understand the effects of these forces. The measurements are taken in the physical space and the knowledge of the geometry of this space is essential to the correct utilization of the observations.

As known from daily experience, the most conspicuous force present on the surface of the earth is gravity. Thus when studying the geometry of the earth geodesists, of necessity, become interested in the earth's gravity field. Consequently, investigation of the geometrical aspects of the gravity field is now being recognized as an integral part of geodesy. Since basic geodesy deals with either stationary or slow moving objects, the gravitational theory needed is that of Newton rather than that of Einstein (see Chapter 1).

This chapter is meant to be a descriptive introduction to the topic, while methods used in the investigations of the gravity field will be given in Part V. Nevertheless, after having digested this chapter, the reader should have gained enough insight into the topic to be able to follow the pertinent arguments of Part IV. Emphasis here is on the terrestrial aspects of the earth's gravity field, meaning that the gravity field on and immediately above the earth's surface is dealt with. The studies of the gravity field in outer space, requiring a somewhat more sophisticated treatment, will be discussed in Part V. Throughout this chapter, the earth is regarded as a rigid body. Whenever necessary, it is pointed out how the non-rigidity of the earth influences our conclusions.

In the first section, the earth's gravity field is defined from the physical and mathematical points of view. The next section is devoted to a description of the magnitude of gravity and how it is handled. Also introduced is the idea of normal gravity and its uses. In the third section, gravity potential is explained, and the terms of equipotential surface and plumb line are defined. The final section deals with the geoid and the deflections of the vertical.

#### 6.1. Gravity field

As we have seen in §1.2, it was ISAAC NEWTON [1687] who first formulated mathematically, in his famous *law of universal gravitation*, the fact that any two

physical bodies attract each other. This law states that a body of mass  $M$  attracts another body of mass  $m$  by a force  $\vec{F}$ , whose magnitude is proportional to the product of the two masses and inversely proportional to the square of their distance  $\Delta r$ :

$$\vec{F} = G \frac{Mm}{\Delta r^2}. \quad (6.1)$$

This force has become known as the *gravitational force* and is also called *gravitational attraction*, or Newton's attraction. The constant of proportionality  $G$ , also denoted in the literature by  $k$ ,  $f$ , or  $\kappa$ , is known as Newton's *gravitational constant*. It can be interpreted as the general property of any mass. Physically, it is the ratio between the behaviour of mass as a source of gravitation and behaviour of the same mass as a responder to gravitation. Its value, determined from various experiments, is  $6.672 \times 10^{-11} \text{ kg}^{-1} \text{ m}^3 \text{ s}^{-2}$  or, equivalently,  $6.672 \times 10^{-8} \text{ g}^{-1} \text{ cm}^3 \text{ s}^{-2}$ , precise to about  $0.001 \times 10^{-8} \text{ g}^{-1} \text{ cm}^3 \text{ s}^{-2}$  [INTERNATIONAL ASTRONOMICAL UNION, 1977].

The gravitational attraction between two bodies is thought to propagate along a straight line with a velocity comparable to the velocity of light. For our purposes, it is adequate to regard the velocity as infinite and thus view the gravitation as having an instantaneous effect at any distance we wish to consider. This is the assumption adopted in classical mechanics.

Taking two physical bodies  $A$  and  $B$ , with masses  $m$  and  $M$ , and considering their dimensions negligibly small compared with their distance, the following vector equation can be written (cf. FIG. 1) for the gravitational force that  $B$  exerts on  $A$ :

$$\vec{F}_{B \rightarrow A} = G \frac{Mm}{|\vec{r}_B - \vec{r}_A|^3} (\vec{r}_B - \vec{r}_A). \quad (6.2)$$

To obtain the force that  $A$  exerts on  $B$ , subscripts  $A$  and  $B$  are interchanged.

What happens if the dimensions of one of the two bodies, say  $B$ , cannot be regarded as negligible? Such would be the case of a small body  $A$  and the earth  $B$ .

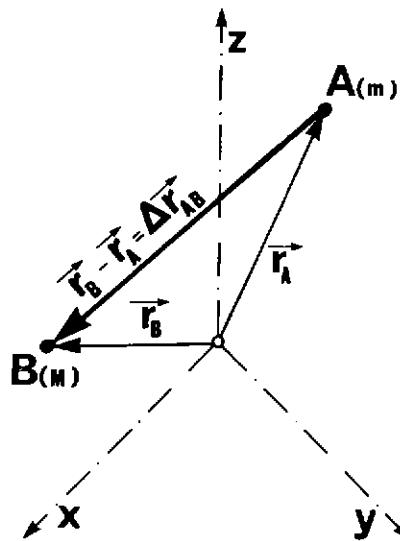


FIG. 6.1. Gravitational attraction between two particles.

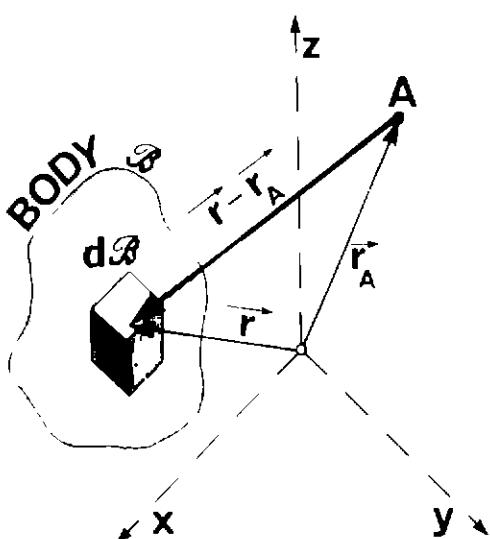


FIG. 6.2. Gravitational attraction of a physical body.

Then the body  $\mathcal{B}$  can be thought of as being composed of a number of small massive elements of volume  $d\mathcal{B}$ , and the attraction of each of these on  $A$  can be investigated separately (FIG. 2). If the independent variable is denoted by  $\bar{r}$ , the mass density within the body by  $\sigma(\bar{r})$ , and  $d\mathcal{B}$  is chosen small enough so that  $\sigma$  in  $d\mathcal{B}$  can be considered constant, then the following relation is obtained:

$$\bar{F}_{d\mathcal{B} \rightarrow A} = G \frac{\sigma(\bar{r}) d\mathcal{B} m}{|\bar{r} - \bar{r}_A|^3} (\bar{r} - \bar{r}_A). \quad (6.3)$$

It has been determined by experiments that gravitational forces are additive [MACMILLAN, 1930]. This means that the sum of forces produced by the elements  $d\mathcal{B}$  is equal to the force exerted by the whole body  $\mathcal{B}$ . By considering the volumes  $d\mathcal{B}$  infinitesimally small, the final equation is obtained by integrating over the body  $\mathcal{B}$  (see §3.2):

$$\bar{F}_{\mathcal{B} \rightarrow A} = \bar{F}(\bar{r}_A) = Gm \iiint_{\mathcal{B}} \frac{\sigma(\bar{r})}{|\bar{r} - \bar{r}_A|^3} (\bar{r} - \bar{r}_A) d\mathcal{B}. \quad (6.4)$$

This equation can be used to study the gravitational force of the earth on bodies the dimensions of which can be considered negligible with respect to the earth. To study the gravitation, though, the density distribution  $\sigma(\bar{r})$  within the earth must be known. But such a distribution is known only approximately. FIG. 3 shows one of the existing density distribution models, according to BULLEN [1963], as obtained from seismic observations. All of the seismic models assume a perfectly spherical distribution, so the density is a function only of a distance from the centre of mass or depth. It can be seen that the gravitational force produced by such an earth model is radial, i.e., the force generated by this body always points towards the centre of

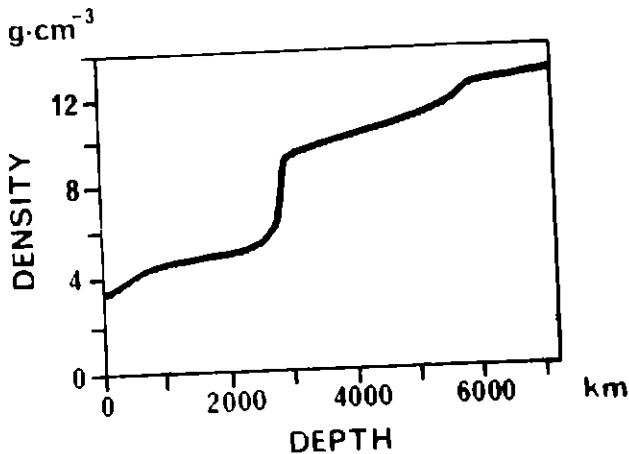


FIG. 6.3. Variation of density with depth.

mass, and its magnitude depends only on the distance from the centre of mass. This means that the gravitation of this model body, on and above its surface, is equivalent to the gravitation of a particle located at the centre of mass, with a mass  $M$ , equal to the mass of the whole body, given as

$$M = \iiint_{\mathcal{B}} \sigma(\bar{r}) d\mathcal{B}. \quad (6.5)$$

But it is already known that the gravitation of such a particle is given by eqn. (2). If the mean radius of the earth ( $R$ ) is taken equal to 6371.009 km, and  $GM$  equal to  $3.986005 \times 10^{20} \text{ cm}^3 \text{ s}^{-2}$  [IAG, 1980], then (2) gives the mean value of the gravitational attraction on the surface of the earth:

$$|\bar{F}| = F \doteq 982.022 [\text{cm s}^{-2}] \times m, \quad (6.6)$$

where  $m$  is the mass of the attracted particle. Since the real distribution of density within the earth is not only radially but also laterally irregular and the earth is not spherical, the gravitational force field is not perfectly radial either. The value of gravitation given by (6) then is only a mean global value.

In the absence of better knowledge of the actual density distribution, eqn. (4) is of limited value to geodesy except to show, theoretically, how gravitation depends on density. Equation (4) also shows that when the density varies with time, so does the gravitational force. This is the case with the real earth, but these variations are minute and thus difficult to detect. Hence in all geodetic work the practice has been to ignore these variations, with the exception of tidal variation (cf. §8.1 and Chapter 25). Here, a stationary density distribution will be assumed.

The fact that the earth is spinning complicates things somewhat even if the earth is assumed to be rigid. The spin of the earth gives rise to an additional force. This force ( $\vec{f}$ ), although only apparent in nature, can be observed to be acting on all earthbound objects (that share the spin with the earth). It is called the *centrifugal force*. Its direction is always perpendicular to the instantaneous spin axis and can be explained as a manifestation of the circular, and therefore accelerated, motion. Its

nature is only apparent because as soon as the object ceases to spin with the earth—ceases to be earthbound—the centrifugal force vanishes.

The magnitude  $f$  of the centrifugal force acting on a particle is known to be equal to [MACMILLAN, 1936]

$$f = p\omega^2 m, \quad (6.7)$$

where  $p$  is the perpendicular distance of the particle from the spin axis,  $\omega$  is the earth's spin angular velocity, and  $m$  is the mass of the particle (cf. FIG. 4). If the spin angular velocity given by eqn. (5.7) is taken,  $\omega = 72.92115 \times 10^{-6} \text{ rad s}^{-1}$ , and  $p = 6378.137 \text{ km}$  [IAG, 1980], then the value of the centrifugal force on the equator is obtained:

$$f \doteq 3.392 [\text{cm s}^{-2}] \times m, \quad (6.8)$$

which is about 0.35% of the gravitational force. On the poles, the centrifugal force vanishes.

The centrifugal force is subject to variations in time in both direction and magnitude. Changes in the magnitude of the spin velocity induce changes in the magnitude of the force; changes in the direction of the spin axis produce changes in the direction of the force. These changes, as seen in §5.3, are very small and can be safely neglected here. They will be discussed, however, in §25.4.

The sum of the gravitational and the centrifugal forces is called the *gravity force*. The field of this force is shown diagrammatically in FIG. 5 by bold arrows. It is easy to understand that the gravity force is stronger on the poles than on the equator. In fact, the difference would be about 0.35% if the earth were spherical. Since the earth

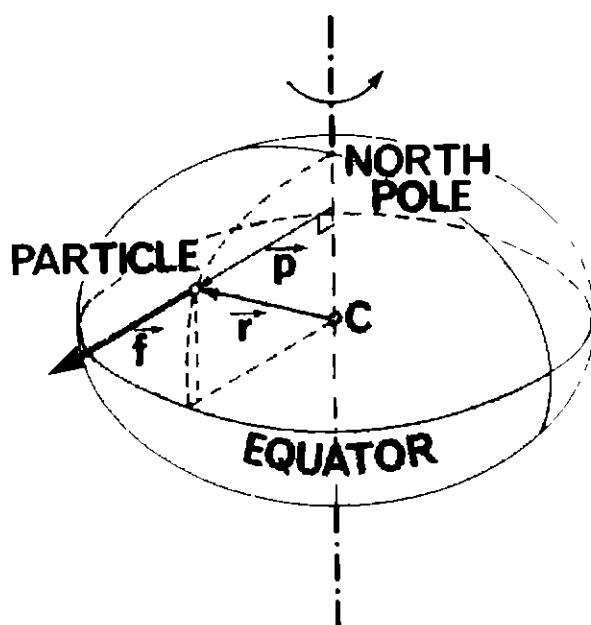


FIG. 6.4. Centrifugal force.

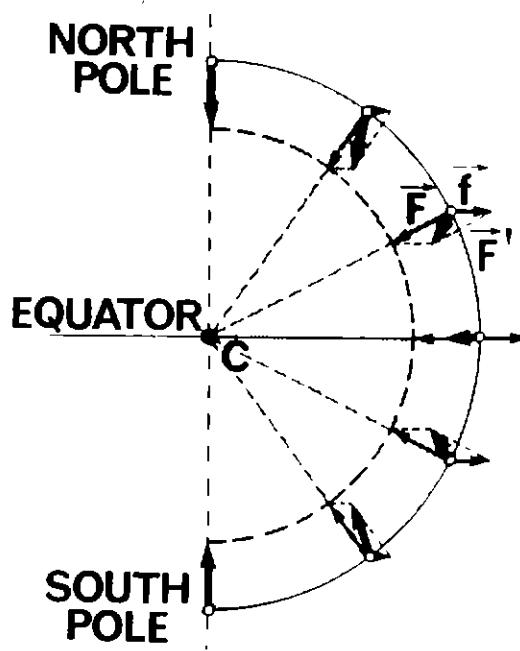


FIG. 6.5. Gravity force.

is oblate, the difference is even more pronounced, being equal to about 0.54%, as will be seen later.

It is usual to work with accelerations rather than forces. To see clearly what is meant by this, let us write the vector equation for the gravity force (summation of eqn. (4) with the vectorial equivalent of eqn. (7)) acting on particle  $A$ ,

$$\begin{aligned}\bar{F}'(\bar{r}_A) &= \bar{F}_{\mathcal{B} \rightarrow A} + \bar{f}_A \\ &= \left\{ G \int \int \int_{\mathcal{B}} \frac{\sigma(\bar{r})}{|\bar{r} - \bar{r}_A|^3} (\bar{r} - \bar{r}_A) d\mathcal{B} + \bar{p}_A \omega^2 \right\} m.\end{aligned}\quad (6.9)$$

It can be seen that the gravity force  $\bar{F}'_A$  is expressed as a product of the term in braces and the mass  $m$  of the particle  $A$ . From Newton's second law, it is known that force is the product of acceleration and mass. Hence the term in braces must be the vector of acceleration. This vector is denoted by  $\bar{g}$  and is called the *gravity vector*; i.e.,

$$\bar{F}'(\bar{r}_A) = \bar{g}(\bar{r}_A)m. \quad (6.10)$$

(Do not confuse the mass  $m$  with metres.) In some publications,  $\bar{g}$  is called the *gravity acceleration vector*.

In a study of the geometrical properties of the gravity force field  $\bar{F}'$ , it is sufficient to concentrate on the acceleration  $\bar{g}$ . The mass  $m$  of the test particle can be regarded as the scale of the field  $\bar{F}'$  (cf. (10)); the *gravity field*  $\bar{g}$  gives the complete geometrical picture of the gravity force field. Note that the gravity (acceleration) field can be

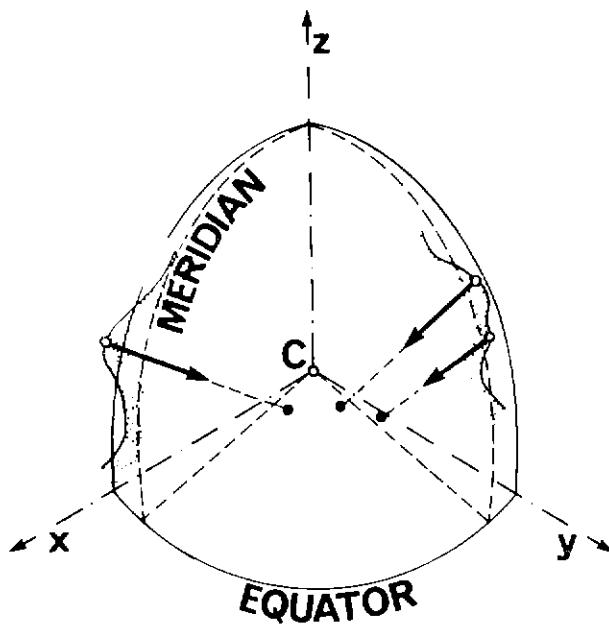


FIG. 6.6. Direction of gravity.

viewed as predicting how the gravity force would act if there was a particle present in the field.

The gravity field, being a vector field, possesses a magnitude (absolute value) and a direction (cf. FIG. 6). The magnitude is easier to deal with because it is a scalar. Its units are gals, named after the famous physicist Galileo Galilei (see §1.2). One gal equals one centimetre per second squared ( $\text{cm}/\text{s}^2$ ); a mean magnitude of gravity on the surface of the earth is of the order of 980.3 Gal (cf. the mean value of gravitation in (6)). The direction of gravity is more awkward to study. It is necessary to introduce a few additional concepts before the directional aspects can be presented (in §6.4).

## 6.2. Gravity anomaly

The magnitude  $g$  of gravity can be measured using any of the existing kinds of gravity measuring instruments (cf. §22.3). A worldwide data bank is maintained by the Bureau Gravimetrique Internationale in Paris, an institution of the IUGG (see §4.2). The several millions of observations so far gathered, from all over the world, show that the magnitudes vary globally and regionally as well as locally. The global range of the variations on the surface of the earth is more than 5 Gal, i.e., more than 0.5% of the average  $g$ . These variations are easily observed even with imprecise instruments: modern instruments measure accurately to within a fraction of a  $\mu\text{Gal}$  ( $1 \mu\text{Gal} = 10^{-6} \text{ Gal}$ ), i.e., to about  $10^{-10} g$ .

These variations have three sources: the different heights of observation points, the oblateness of the earth, and the uneven lateral distribution of masses within the earth. How are these irregularities of the gravity field portrayed? There are two different concepts used here—one for depicting the irregularities in the space above

the earth (spatial variations), and another used for the surface of the earth (terrestrial variations). As stated earlier, the former will not be treated in this chapter, only the latter will.

Let us first examine the variations resulting from the first source.

(a) To investigate the *gravity variations with height*, the usual starting point is the first approximation to (the magnitude of) gravity given, according to (1), by

$$g \doteq G \frac{M}{r^2}, \quad (6.11)$$

where  $r$  is the distance from the earth's centre of mass. Then a straightforward differentiation with respect to  $r$  gives the gravity gradient in radial direction:

$$\frac{dg}{dr} \doteq -2 \frac{GM}{r^3}. \quad (6.12)$$

With the realization that the increment  $dr$  in radial distance is very nearly the same as the increment  $dH$  in height, the following expression for the increment of gravity with height is obtained:

$$dg \doteq -2 \frac{GM}{r^3} dH. \quad (6.13)$$

Substituting the proper value for  $GM$  and taking  $r$  to be equal to the mean radius of the earth, the approximate expression for the gravity increment with height at or near the surface of the earth is finally obtained:

$$dg \doteq -0.308 [\text{mGal m}^{-1}] dH. \quad (6.14)$$

First, note that  $dg$  is negative for a positive  $dH$ , which means that the gravity magnitude decreases with increasing height; this can also be verified directly from (11). Second, we see that  $g$  decreases by only 1% with an increase in altitude of about 32 km, i.e., the gravity decreases only by about 0.28% in a climb to the top of Mt. Everest.

The correction to gravity for the height effect given by (14) is called the *free air correction*. In Chapter 21, other techniques for evaluating the height effect will be considered. Whichever correction for height is used, however, the corrected gravity still varies globally with latitude (effect of oblateness), as well as regionally and locally (effect of irregular mass distribution).

(b) Before talking about the *gravity variation due to the oblateness of the earth*, the following should be pointed out: all the irregularities, although significant and easily observed, are still minute compared with the magnitude of gravity itself. Therefore, an attempt is made to express the bulk of the magnitude by an analytical expression, and subtract it from the actually observed values. The most widely used technique is to first correct the observed gravity for the height effect, as previously stated. The gravity, corrected in the above described manner, is then compared with an analytically defined *reference gravity*. It is this difference that is taken as the measure

of variation in the magnitude. For geodetic purposes, the reference gravity can be selected arbitrarily with the one aim only in mind—to keep the average difference from actual gravity as small as possible.

Evidently, the reference gravity could be taken as given by (11); i.e., we could refer the actual gravity to a radial field. By doing this, all the magnitudes would be reduced, but the main variation would prevail; the differences of the order of 5 Gal, due to the earth's oblateness, would still remain. These differences can be reduced by one order of magnitude by letting the reference gravity also reflect the oblateness of the earth. This can be done by defining a 'massive' biaxial ellipsoid (ellipsoid of revolution), concentric with the earth (geocentric), the minor axis of which coincides with the polar principal axis of inertia of the earth. An analytical expression for the hypothetical gravity field generated by this ellipsoid, assuming that it spins around its minor axis with the same angular velocity as that of the earth and has a rotationally symmetrical mass distribution, can then be derived (for a detailed treatment, see §20.3). Such a reference gravity field is called the *normal gravity field* and is represented by the *normal gravity vector* denoted by  $\bar{\gamma}$ . The normal field is a function of both the distance from the centre of mass of the earth and latitude  $\phi$ ; it is, however, rotationally symmetrical, i.e.,  $\bar{\gamma}$  does not depend upon longitude. It is customary to express  $\bar{\gamma}$  as a function of latitude and height  $h$  above the geocentric ellipsoid.

BOWIE AND AVERS [1914] made one of the first attempts to define a normal gravity field that would absorb the effect of the earth's oblateness. Their formula for the magnitude of normal gravity is still used today for some geodetic work. It reads:

$$\begin{aligned}\gamma = & 980.624(1 - 0.002644 \cos 2\phi + 0.000007 \cos^2 2\phi) \\ & - 0.3086 h - 0.0002 h \cos 2\phi + 7.1 \times 10^{-8} h^2 \text{ Gal},\end{aligned}\quad (6.15)$$

where  $h$  is the height in metres.

As stated earlier, it is usual to compare the reduced gravity—corrected for the height effect—with the normal gravity. For this purpose, the normal gravity on the surface of the geocentric ellipsoid, generally denoted by  $\gamma_0$ , is used. The Bowie-Avers normal gravity  $\gamma_0$  is given by eqn. (15) by putting  $h=0$ :

$$\gamma_0 = 980.624(1 - 0.002644 \cos 2\phi + 0.000007 \cos^2 2\phi) \text{ Gal.} \quad (6.16)$$

Other researchers have proposed different formulae.

In order to unify the definition of normal gravity throughout the world, the IAG (see §4.2) adopted, in its plenary session in Stockholm in 1930, a formula for normal gravity [CASSINIS, 1930],

$$\gamma_0 = 978.0490(1 + 0.0052884 \sin^2 \phi - 0.0000059 \sin^2 2\phi) \text{ Gal,} \quad (6.17)$$

and recommended it to its member countries for use in all gravity work. The formula subsequently became known as the *International Gravity Formula 1930*. In 1967, the IAG General Assembly approved new parameters for the geocentric biaxial ellipsoid. The normal gravity on this international ellipsoid was distributed according

to the following approximate formula [IAG, 1971]:

$$\gamma_0 \doteq 978.031\,85(1 + 0.005\,278\,895 \sin^2\phi + 0.000\,023\,462 \sin^4\phi) \text{ Gal}, \quad (6.18)$$

with the maximum error of  $4 \mu\text{Gal}$ . Equation (18) was called the *International Gravity Formula 1967*. The newest *International Gravity Formula 1980* [IAG, 1980] was adopted by the IAG General Assembly in Canberra. This formula, recommended to be used by the IAG member countries, reads, to an accuracy of  $0.7 \mu\text{Gal}$ :

$$\begin{aligned} \gamma_0 \doteq 978.032\,7 &(1 + 0.005\,279\,041\,4 \sin^2\phi \\ &+ 0.000\,023\,271\,8 \sin^4\phi + 0.000\,000\,126\,2 \sin^6\phi) \text{ Gal}. \end{aligned} \quad (6.19)$$

The difference between the reduced actual gravity and normal gravity on the ellipsoid is called the *gravity anomaly* and is denoted by  $\Delta g$ . According to the way actual gravity is corrected for the height effect, there are different kinds of gravity anomalies. The use of the free air correction distinguishes the *free-air gravity anomaly*. Other kinds of anomalies will be seen in Chapter 21.

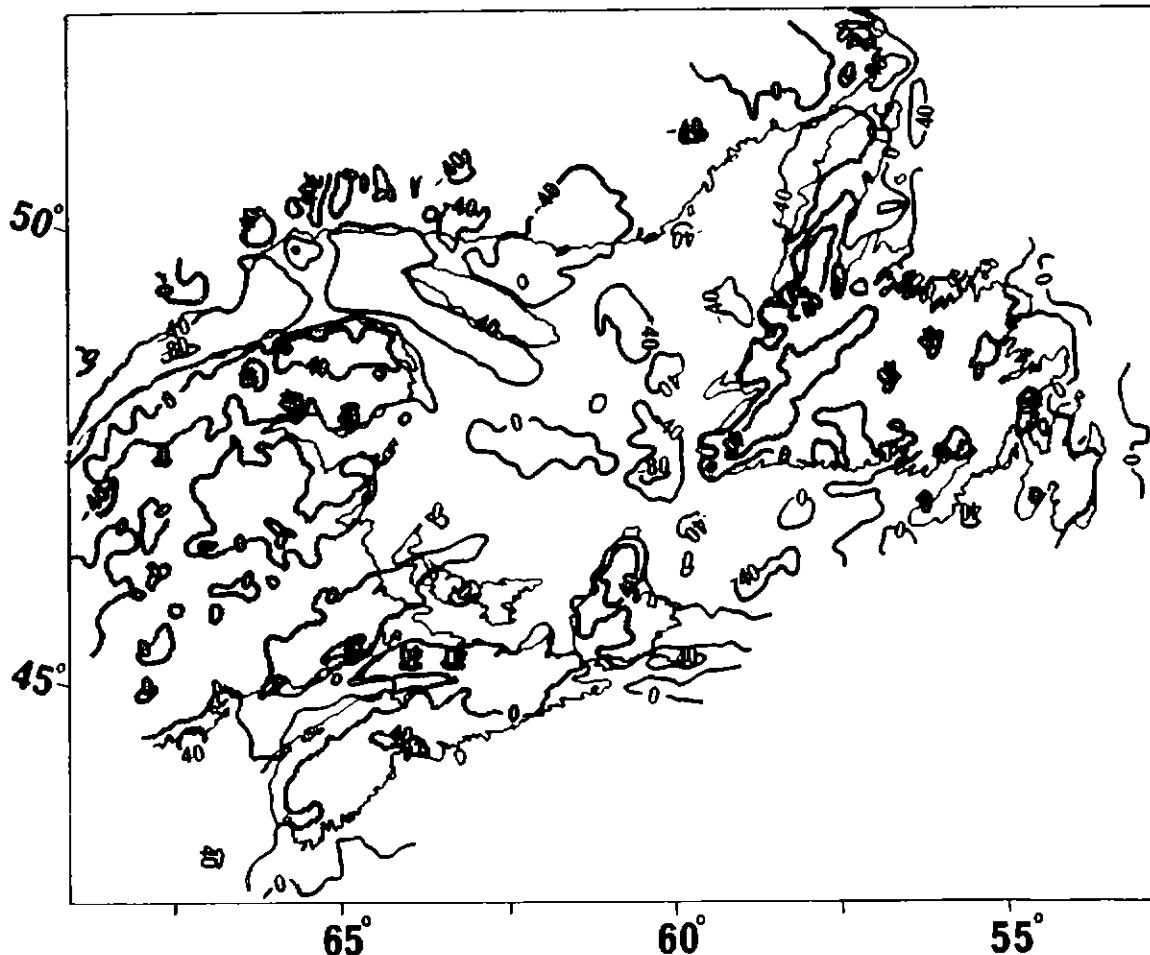


FIG. 6.7. Free air gravity anomaly referred to the 1967 formula in Eastern Canada. Contours in milligals. (Courtesy of the Earth Physics Branch, DEPARTMENT OF ENERGY, MINES AND RESOURCES [1977a], Ottawa, Canada.)

According to which normal gravity formula is used for computing the gravity anomaly, gravity anomalies are referred to the 1930 formula, the 1967 formula, etc. A regional map of free air gravity anomalies in Eastern Canada, referred to the 1967 formula, is shown in FIG. 7. FIG. 8 depicts free air anomalies, based on the 1930 formula, for the whole of Canada [NAGY, 1973]. Notice that apart from a few small areas, where the anomaly reaches 100 mGal (0.01%  $g$ ), the variation is fairly small. The large variation of more than 5000 mGal, caused by the earth's oblateness, has been absorbed by the normal gravity. This point is even better illustrated by the global map of gravity anomalies, derived from satellite observations [GAPOSHKIN, 1973], as given in FIG. 9. On this map, any global variations caused by the earth's oblateness, if they were present, would be clearly indicated. It can be seen, however, that on all three maps the local and regional irregularities still persist. In the global solution the local irregularities are smoothed out, so they appear smaller than they actually are.

(c) Whatever remains—after removing the height and oblateness effects—are variations of gravity due to the irregular distribution of masses within the earth. It is this relationship that makes gravity observations of value to other earth sciences. A positive gravity anomaly ( $g > \gamma$ ) shows that there are relatively denser masses

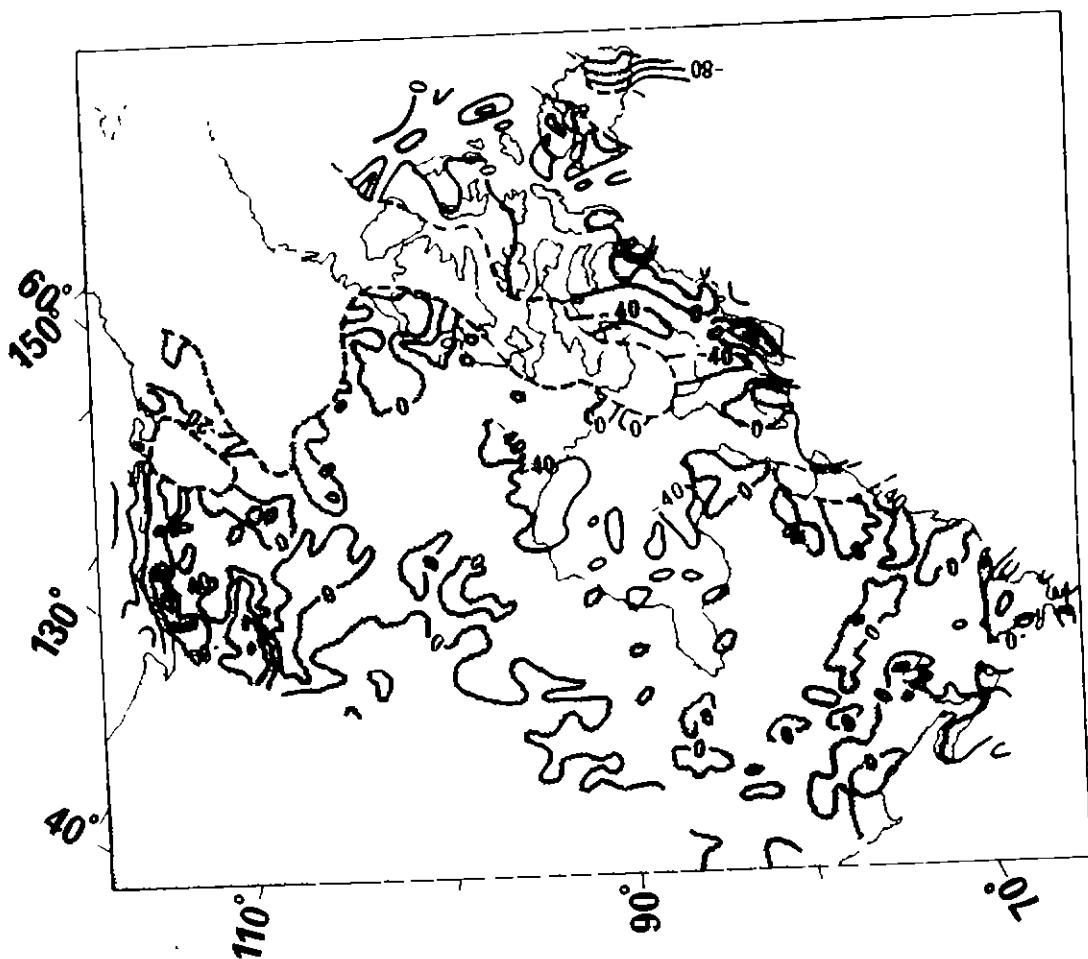


FIG. 6.8. Free air gravity anomaly referred to the 1930 formula in Canada. Contours in milligals.

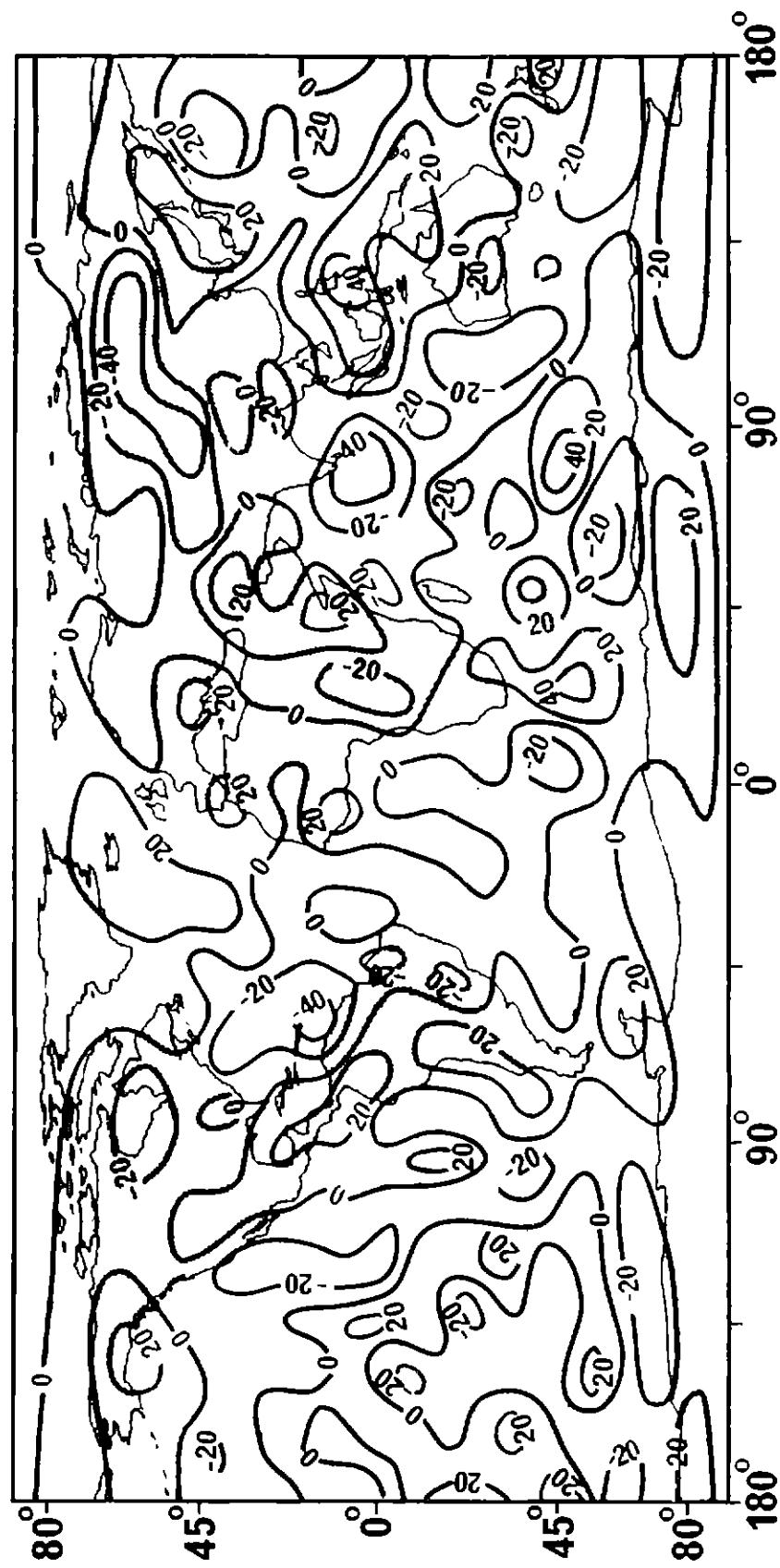


FIG. 6.9. Global free air gravity anomaly referred to the 1967 formula. Contours in milligals.

underneath; i.e., that there is a positive density anomaly underground. A negative gravity anomaly ( $g < \gamma$ ) corresponds to a negative density anomaly. The finding of the possible density distribution corresponding to observed gravity variations is known as gravity interpretation, not to be dealt with in this book. One rule, however, can be formulated: The broader features of the gravity anomalies reflect density anomalies in greater depths; smaller features are due to the shape of the earth's surface and near surface density anomalies [GARLAND, 1965, PICK ET AL., 1973].

As stated in §6.1, using the magnitude of gravity in terms of anomalies is one possible way of mapping the earth's gravity field. Another possibility is to use gravity direction instead of magnitude. Before explaining this approach, the concept of potential must first be introduced.

### 6.3. Gravity potential

The gravity field being a vector field means that there is a vector, i.e., a triplet of numbers, assigned to every point in space. It is much more expedient to work with a scalar field, where there is just one number needed at every point. A question now arises: Is it possible to represent completely the vector field by a scalar field? The answer is yes, at least for some vector fields including the earth's gravity field.

Let us take an arbitrary, closed, spatial curve  $\mathcal{C}$  within the vector field  $\bar{v}$ . If the following equation,

$$\oint_{\mathcal{C}} \bar{v}(\bar{r}) \cdot d\bar{r} = 0, \quad (6.20)$$

where  $d\bar{r}$  is directed along the curve  $\mathcal{C}$  (see §3.2), holds for any curve  $\mathcal{C}$  (cf. FIG. 10), then the field  $\bar{v}$  is called *irrotational*. If a field is irrotational, then there exists a scalar field  $V$  such that

$$\bar{v}(\bar{r}) = \nabla V(\bar{r}) = \text{grad } V(\bar{r}). \quad (6.21)$$

If time is considered, then for the above equation to hold, the field  $\bar{v}$  also has to be *conservative*, i.e., not changing with time. This scalar field  $V$  is called the potential

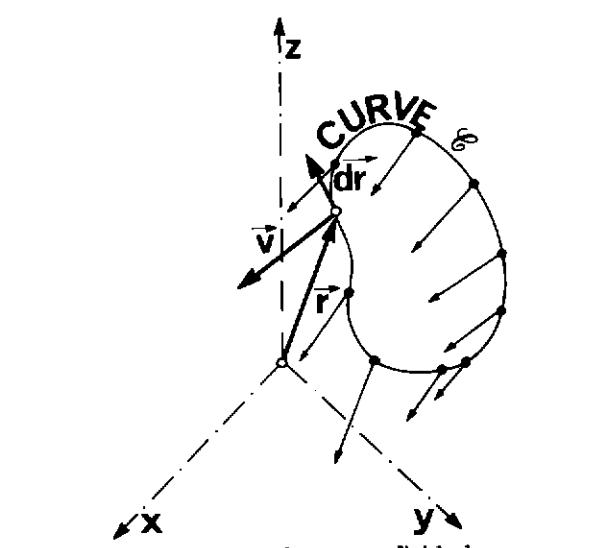


FIG. 6.10. Integration of a vector field along a curve.

energy of  $\bar{v}$ , and  $\bar{v}$  is then the gradient of  $V$  (see §3.2). From the physical point of view,  $l$  is the amount of work needed to overcome the force  $\bar{v}$ . Its physical units are  $\text{g cm}^2 \text{ s}^{-2}$ .

It has been shown that the gravity force field is irrotational [MACMILLAN, 1930], and as such, it has a potential energy corresponding to it. Furthermore, because the gravity (acceleration) field  $\bar{g}$  differs from the gravity force field only by a scale  $m$  (cf. (10)), it is easy to see, from (21), that the gravity field can be expressed as

$$\bar{F}' = m\bar{g} = \nabla V = m\nabla W. \quad (6.22)$$

In other words, there also exists a scalar field  $W$  such that

$$\boxed{\bar{g} = \nabla W.} \quad (6.23)$$

This scalar field is known as the *gravity potential*.

The gravity potential  $W$  is sometimes treated as the negative amount of work needed to overcome the gravity force  $m\bar{g}$  acting on a unit mass  $m$ . But its physical units are  $\text{cm}^2 \text{ s}^{-2}$  and thus do not reflect the presence of any mass. It is preferable, therefore, to view  $W$  as ‘work’ in the kinematic sense, i.e., involving no mass. Since the potential  $W$  differs from the potential energy  $V$  only by scale  $m$ —the mass of the attracted particle—the geometries of the  $V$  and  $W$  scalar fields are the same.

Let us have another look at (9): it can be seen that the gravity acceleration can be expressed as a sum of the triple integral, representing the gravitational acceleration, and another term representing the centrifugal acceleration. Since the (gradient) differential operator  $\nabla$  is a linear operator, the gravity potential  $W$  can also be expressed as the sum of a *gravitational potential*  $W_g$  and a *centrifugal potential*  $W_c$ . If the gravitational acceleration is denoted by  $\bar{g}_g$  and the centrifugal acceleration by  $\bar{g}_c$ , then (see §3.2)

$$\bar{g} = \bar{g}_g + \bar{g}_c = \nabla W_g + \nabla W_c = \nabla(W_g + W_c). \quad (6.24)$$

The reader can verify, by evaluating the gradients and comparing them with (9), that the two potentials are given by the following formulae:

$$\boxed{W_g(\bar{r}_A) = G \iiint_{\mathcal{B}} \frac{\sigma(\bar{r})}{|\bar{r} - \bar{r}_A|} d\mathcal{B}.} \quad (6.25)$$

$$\boxed{W_c(\bar{r}_A) = \frac{1}{2} p_A^2 \omega^2.} \quad (6.26)$$

After inspecting these formulae, one is lead to the realization that while  $W_g$  decreases above the earth (i.e., for  $|\bar{r}_A| > |\bar{r}|$ ), being inversely proportional to the distance  $|\bar{r} - \bar{r}_A|$ ,  $W_c$  increases proportionally to the square of the distance  $p_A$  from the spin axis. For instance, taking a direction from the centre of mass in the equatorial plane, it is found that the potential changes with distance, as shown in FIG. 11. When interpreting this diagram, one has to bear in mind that the combined potential  $W$  acts only on earthbound bodies or particles. As soon as the body stops

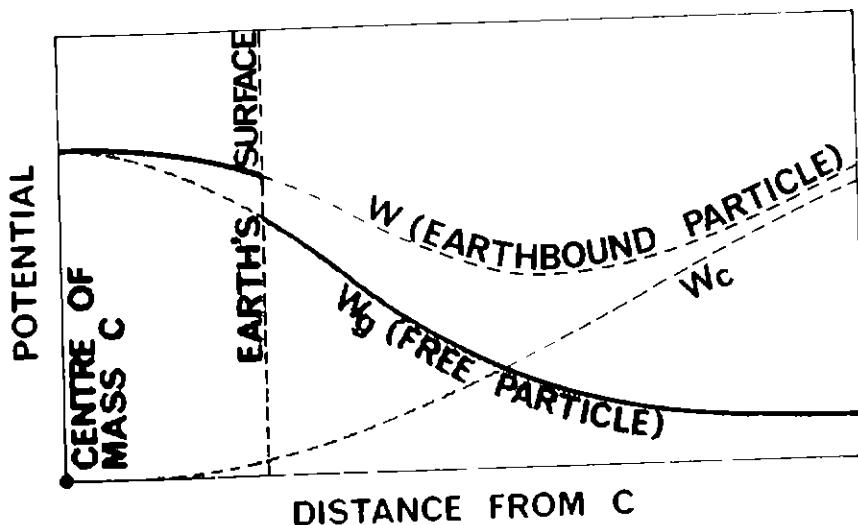


FIG. 6.11. Gravity and gravitational potentials.

spinning with the earth, the centrifugal potential  $W_c$  ceases to be relevant, and only the gravitational potential  $W_g$  prevails.

The gravity potential  $W$  must contain all the information there is about the gravity field. Hence it is to be expected that a 'smooth' potential corresponds to the smooth gravity field and an 'irregular' potential to the irregular gravity field. How can the potential then be used to depict the irregularities seen in the gravity field? The simplest way to use the gravity potential  $W$  to characterize these irregularities, is to use its equipotential surfaces and its lines of force. The *gravity equipotential surface* is a surface on which the gravity potential is constant. The general equation of an equipotential surface is

$$W(\vec{r}) = \text{const.} \quad (6.27)$$

Obviously, an infinite number of equipotential surfaces can be found just by assigning different values to the potential. The lines of force are the curves to which the gradient of the potential, i.e., the gravity field itself, is tangent at every point. The lines of force of the earth's gravity field are called the *plumb lines* (cf. FIG. 12).

There are several properties of the earth's equipotential surfaces of importance to geodesy.

- (a) To begin with, the equipotential surfaces never cross each other: they are closed surfaces, completely enveloping each other like the layers of an onion.
- (b) They are also continuous, without breaks.
- (c) They do not possess any steps or sharp edges, i.e., tangent planes at any two infinitesimally close points on any one equipotential surface are only infinitesimally different.
- (d) The local radii of curvature of the equipotential surfaces vary smoothly from point to point, with the exception of points where the mass density changes suddenly. An example of such singular points are points on the surface of the earth.

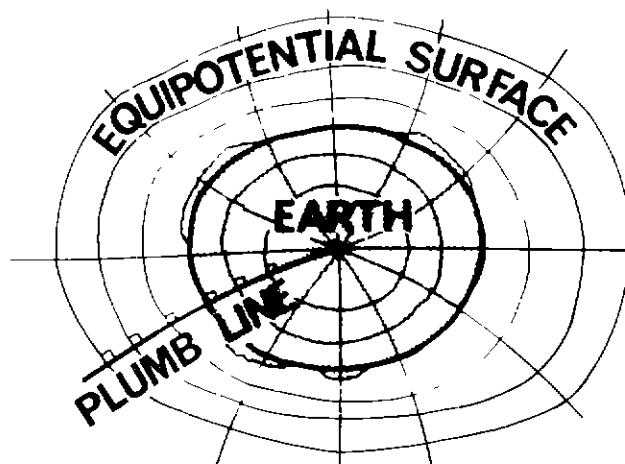


FIG. 6.12. Equipotential surfaces and plumb lines of the earth's gravity field.

(e) Finally, the equipotential surfaces are convex everywhere. This means that they do not have any lows, troughs, or valleys. (For didactic reasons, this property is not always adhered to in our illustrations.) For more details, and precise mathematical treatment, the reader is referred to MACMILLAN [1930] and LANDKOF [1972].

When moving along an equipotential surface, no change in the potential is experienced and thus no work, in the static sense, is done, either positive or negative. Therefore this movement cannot go with or against the direction of the force field. The consequence is that the force lines must all be perpendicular to the equipotential surfaces. Since the direction of the plumb line is commonly referred to as the *vertical direction*, the equipotential surfaces define the horizontal direction; thus they are also called *level surfaces*. Contrary to popular belief, a massive thread of a plumb bob does not follow a plumb line, nor does the trajectory of a free falling mass. The reader is left to reason these statements out.

If a cross section of an equipotential surface is drawn (i.e., a curve orthogonal to the gravity vectors in FIG. 5), it can be seen that it makes an oblate curve. All the equipotential surfaces make an oblate spatial pattern (cf. FIG. 12) reminiscent of a series of concentric ellipsoids. Due to the irregular density distribution however, the equipotential surfaces are also somewhat irregular. Because their radii of curvature in various directions change irregularly from point to point, the plumb lines are also curved in all directions—they possess not only a curvature (bend) but also a torsion (twist), and are thus spatial curves. It is worth remembering that all these irregularities, although significant, are relatively small.

It has been seen that there is a definite relation between the equipotential surfaces and the direction of gravity—they are mutually perpendicular. The question often asked is: What is the relation between the equipotential surfaces and the magnitude of gravity? It is the spacing of the equipotential surfaces that is directly related to the magnitude of gravity. The closer together the surfaces, the stronger the gravity field (the larger the  $g$ ), and vice versa (cf. FIG. 13). This becomes obvious once it is realized that  $g$  is merely the difference of the potentials of two infinitesimally close

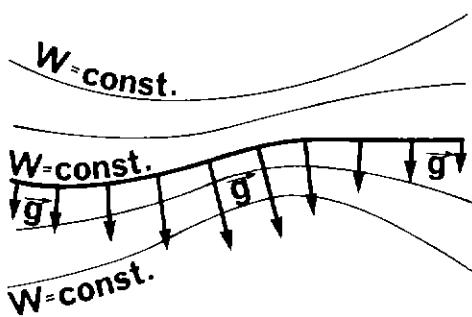


FIG. 6.13. Gravity on an equipotential surface.

equipotential surfaces divided by their separation, i.e.,

$$g = |\nabla W| \doteq -\frac{\partial W}{\partial h}. \quad (6.28)$$

The above reasoning also gives the answer to one often posed and often wrongly answered question: Is gravity on an equipotential surface constant? A look at FIG. 13 convinces us that, generally speaking, the (magnitude of) gravity on an equipotential surface varies.

Another feature of FIG. 12, the fact that, globally speaking, the equipotential surfaces converge more on the poles, can now be explained: this is a consequence of gravity being stronger at the poles than at the equator. Taking, for instance, the International Gravity Formula 1967 (eqn. (18)) as adequately depicting the global gravity field in this context, we get the difference of the two gravity values  $g_E$  (for  $\phi = 0^\circ$ ) and  $g_P$  (for  $\phi = 90^\circ$ ) being approximately equal to 5 Gal, or more precisely,

$$g_P - g_E \doteq 5.186 \text{ Gal} \doteq 5.3 \times 10^{-3} g_E. \quad (6.29)$$

If (28) is rewritten as

$$H_P g_P = H_E g_E = \delta W = \text{const.}, \quad (6.30)$$

where  $H$  is the height of a selected equipotential surface above the geoid, then the height at the poles ( $H_P$ ) equals  $1 - 5.4 \times 10^{-3} = 0.9946$  multiplied by the height at the equator ( $H_E$ ). The convergence is thus 0.54 percent. Another way of viewing this property of the earth's gravity field is that more work is needed to lift a body of constant mass at the pole than it is at the equator. The reasoning behind this statement is left to the reader.

It should be clear by now that a surface of a homogeneous fluid in equilibrium coincides with the pertinent patch of one of the earth's equipotential surfaces. Suppose that the fluid surface differs from the equipotential surface; then there would be differences in potential along the fluid surface, or, in other words, there would be a component of gravity force acting along the fluid surface. The differences in potential, or equivalently, the component of gravity force tangent to the surface, would then create a flow that would again bring the fluid surface into equilibrium, i.e., into coincidence with the equipotential surface.

This is what is happening on the earth's surface locally as well as globally. The surfaces of lakes and oceans tend to follow the gravity equipotential surfaces with only minor deviations due to both the lack of homogeneity in water and the external influences (cf. Chapter 7). Even the earth as a whole behaves under permanent stresses as a viscous body, i.e., as if it were composed of a highly viscous fluid. Thus, it too tends to adjust itself to an equipotential shape. More will be said about this in the next chapter.

#### 6.4. Geoid and deflections of the vertical

The one gravity equipotential surface of particular interest is that which best approximates the (mean) sea level over the whole earth. It is called the *geoid*. Gauss (see §1.3) described the geoid as being the mathematical figure of the earth, and as such it is a key surface in geodesy, playing a fundamental role in positioning, as will be shown in Part IV.

In the first approximation, i.e., up to a few metres, the geoid is represented by the mean sea level. It generally passes underneath the continents at a depth equal to the height of the terrain above the sea level. The geoid, of course, possesses all the properties (§6.3) ascribed to an equipotential surface.

Observations have shown that the geoid can also be approximated—up to some tens of metres—by a biaxial geocentric ellipsoid whose minor axis coincides with the earth's principal polar axis of inertia. It then becomes natural to use a concept similar to the one used for normal gravity, i.e., to accept an analytically defined '*normal body of the earth*' in terms of the best fitting, geocentric biaxial ellipsoid called, by some authors, the *mean earth ellipsoid* (see FIG. 14). Usually, the same ellipsoid that generates normal gravity (cf. §6.2) is used to refer the geoid to; in this

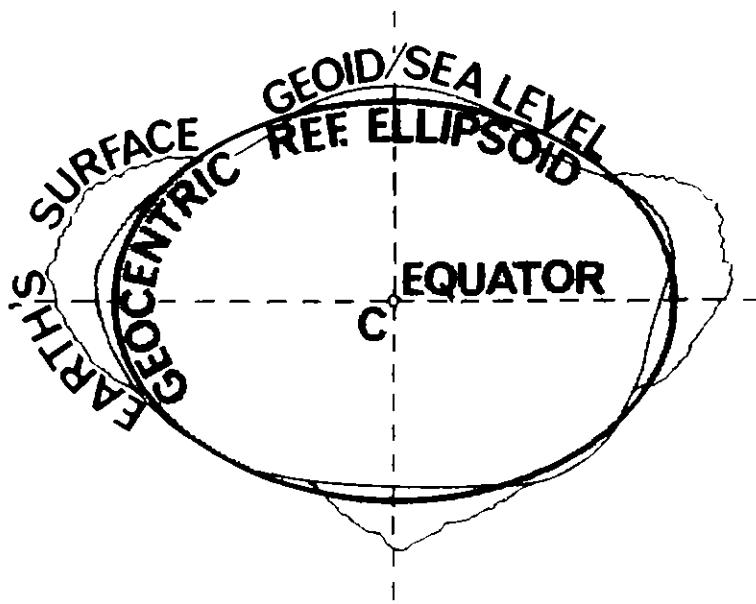


FIG. 6.14. Biaxial ellipsoid as a normal body of the earth.

context it is also spoken of as the *geocentric reference ellipsoid*. More will be said on this topic in Chapter 7.

It is useful to think about the geocentric reference ellipsoid as generating not only the normal gravity but also the potential corresponding to normal gravity. Denoting this *normal potential* by  $U$  results, once more, in the known relation

$$\vec{\gamma} = \nabla U. \quad (6.31)$$

To make the parallel between the actual and normal gravity fields even closer, it is required that the normal potential on the surface of the reference ellipsoid be constant and as close to the actual potential on the geoid as possible. The surface of the ellipsoid thus becomes one of the equipotential surfaces of the normal gravity field which some authors speak of as an equipotential ellipsoid. The reason for striving for this parallel is to make various investigations and computations easier. The 1967 international geocentric reference ellipsoid is assumed to have a potential of  $6.263\,686\,085 \times 10^{11} \text{ cm}^2 \text{ s}^{-2} = 6.263\,686\,085 \times 10^6 \text{ kGal m}$  [IAG, 1980].

The definition of the normal gravity field described above leads to paired quantities: actual gravity field–normal gravity field; actual equipotential surface–normal equipotential surface; actual plumb line–normal plumb line; geoid–geocentric ellipsoid; etc. In the remainder of this section, two such pairs will be treated briefly; geoid–ellipsoid and directions of actual and normal gravity vectors.

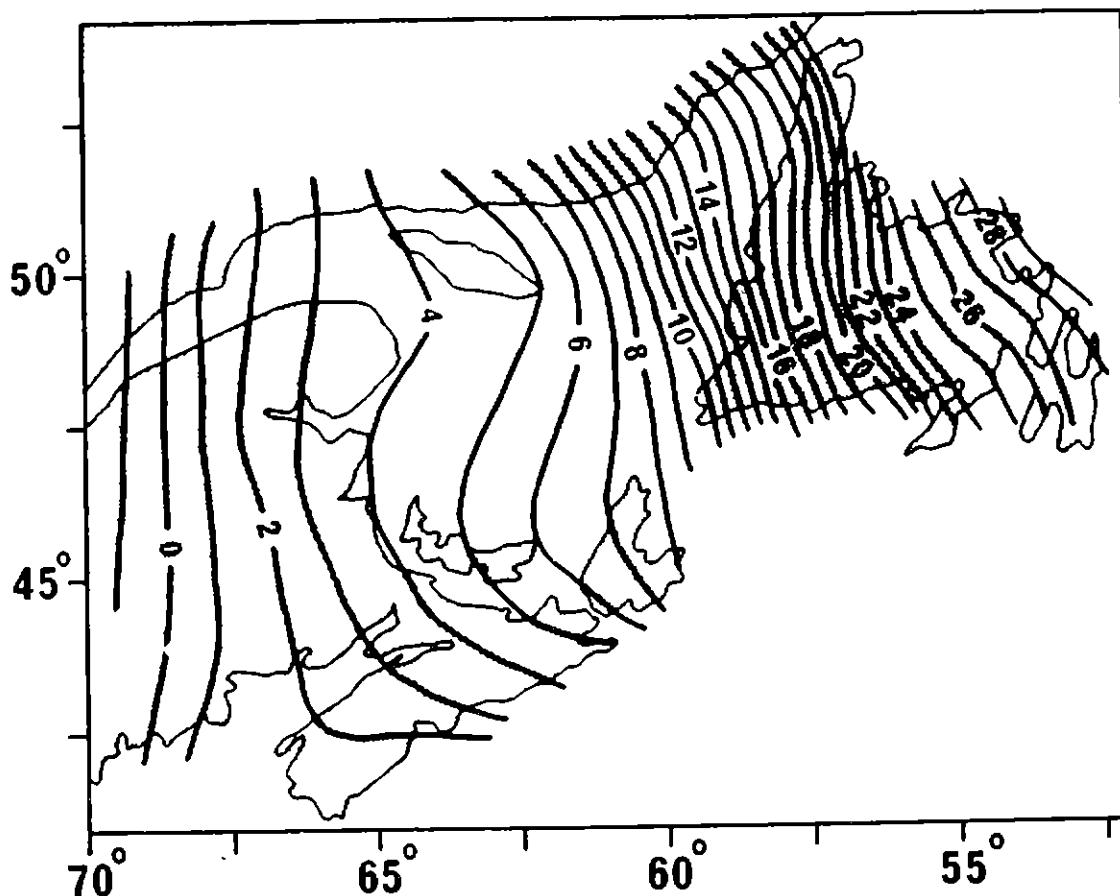
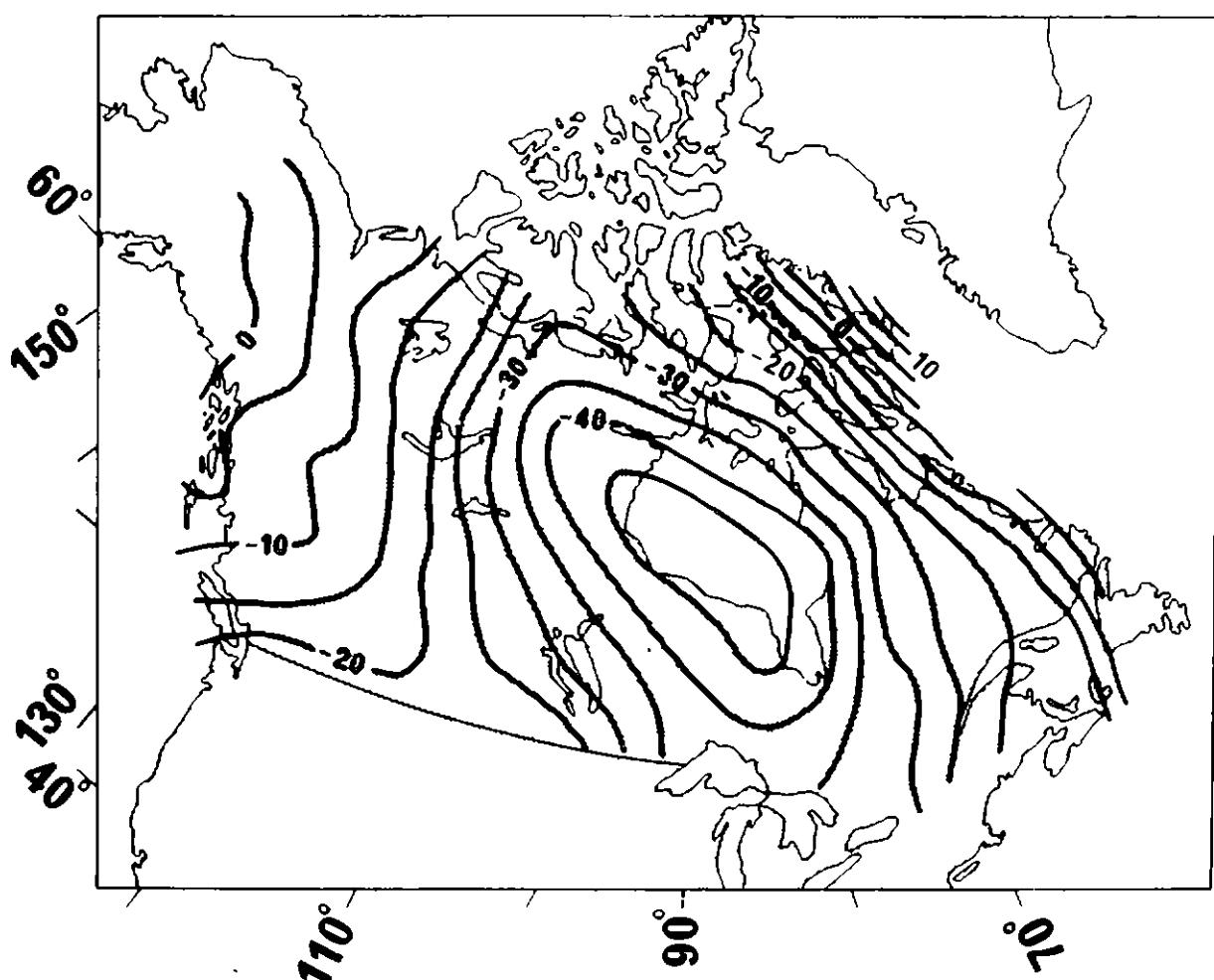


FIG. 6.15. Relative geoid in Eastern Canada referred to the NAD 27. Contours in metres.

Separation between the geocentric reference ellipsoid and the geoid is called the *geoidal height*, or *geoidal undulation*, and is generally denoted by  $N$ . This is sometimes called *absolute geoidal height* because it relates the geoid to an ‘absolute’, i.e., a geocentric reference ellipsoid. There also exists *relative geoidal height*, which refers the geoid to another kind of reference ellipsoid that is not geocentrically located. Relative geoidal heights will be discussed in Parts IV and V. Let it suffice here to show an example—see FIG. 15—of a relative geoid for Eastern Canada [MERRY AND VANÍČEK, 1974] related to a non-geocentric reference ellipsoid called the (North American Datum 1927) NAD 27 (cf. §7.3).

An example of an absolute geoid for the whole of Canada [VINCENT ET AL., 1972] is given in FIG. 16. This particular solution is referred to a geocentric ellipsoid, 34 m smaller in size than the 1967 international ellipsoid (the difference in shapes being insignificant), which also has a suitably smaller mass. Therefore, the geoidal heights referred to these two ellipsoids should be directly comparable. More about this solution will be said in §22.1.

Similar to the case of anomalies, it can be said that the broader features of the geoid are due to deep-seated density anomalies, while short wavelength features reflect the near surface density distribution. Hence for global studies, global solutions are used. FIG. 17 shows one of the many existing global solutions [RAPP, 1974]



## EARTH AND ITS GRAVITY FIELD

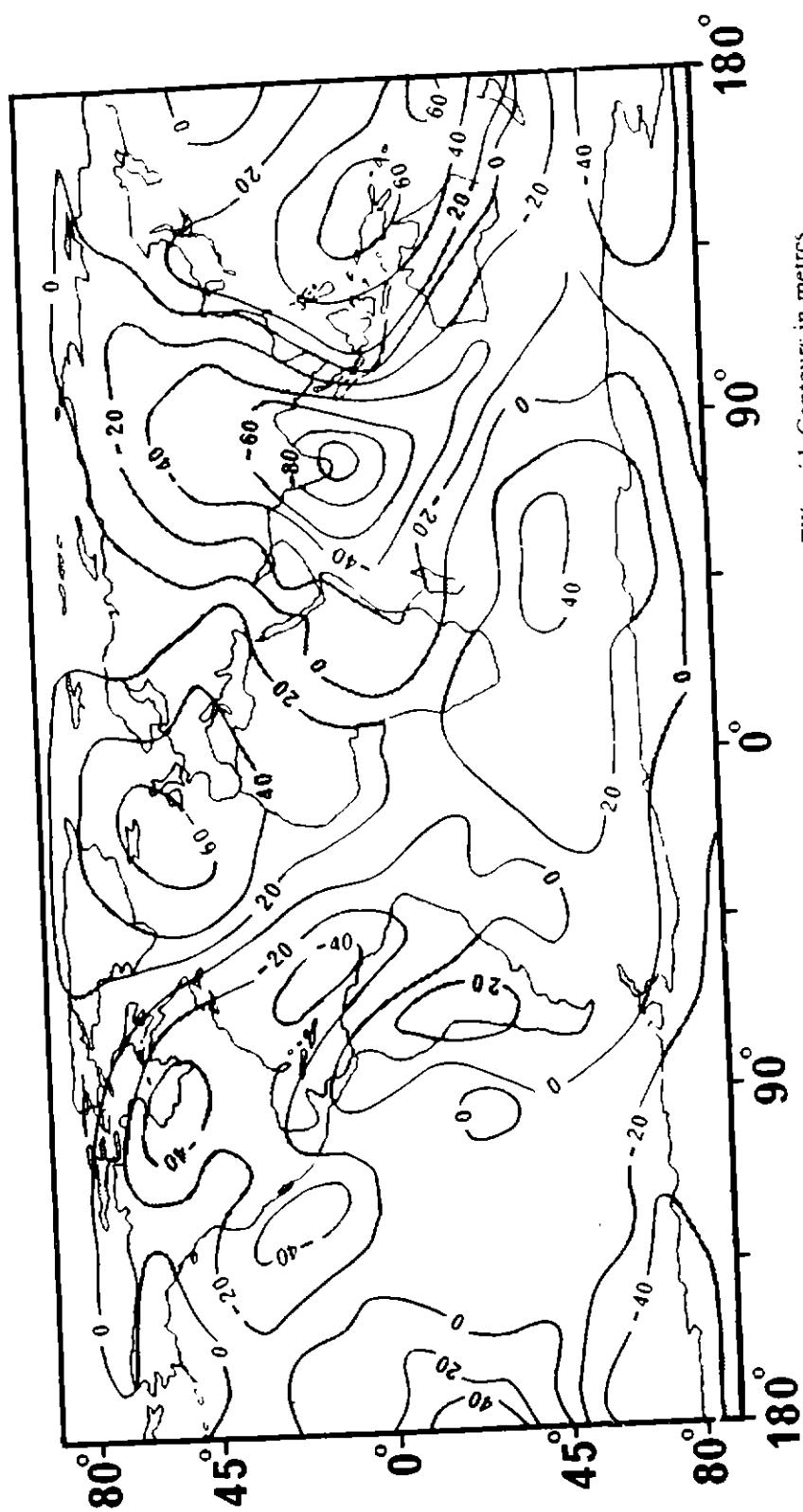


FIG. 6.17. Global geoid referred to the 1967 International Reference Ellipsoid. Contours in metres.

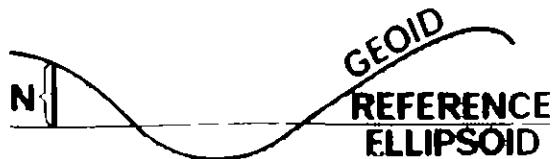


FIG. 6.18. Geoidal profile.

for the geoid. This solution is only approximate in that just the long wavelength features of the geoid are indicated. Consequently, the geoid looks much smoother than it actually is. The various techniques to compute the geoid will be given in Part V.

One must bear in mind that, in reality, the lows on the maps are not concave. As stated earlier, the geoid is a convex surface; these apparent regions of concavity are caused by the 'straightening' of the ellipsoid onto a plane. Note that the largest geoidal height (in absolute value) occurs just south of India, where the geoid passes about 100 m below the reference ellipsoid.

A very interesting relation between the amplitudes and wavelengths of geoidal features was detected empirically by KAULA [1966a]. This relation is best seen on a geoidal profile, such as that shown in FIG. 18. When  $N$  is developed into a trigonometrical Fourier series (see §3.2), then, on average (i.e., for many profiles at various locations and in various directions), the amplitudes  $A_n$  (Fourier's coefficients) diminish with increasing wave number  $n$  — for a detailed discussion see §20.2. If  $n = 1$  is taken as corresponding to the wavelength of 40 000 km (once around the earth), then Kaula's rule of thumb reads

$$A_n \doteq R / (n^2 10^5), \quad (6.32)$$

where  $R$  is, once more, the mean radius of the earth. For example, if a feature is 2000 km long (corresponding to the wavelength 4000 km, or  $n = 10$ ), then its amplitude may be expected to be, on average, about 64 centimetres. Subsequent investigations have confirmed this rule of thumb to be valid for features longer than about 500 km (e.g., BROWN ET AL. [1972]).

As pledged at the end of §6.1, it can now be shown how the directional irregularities of the earth's gravity field are treated. To show these irregularities, the directions of both the actual and the normal gravity vectors are used. FIG. 19 depicts the two gravity vectors  $\bar{g}_0$  and  $\bar{y}_0$  on the geoid and geocentric ellipsoid respectively. The spatial angle between  $\bar{y}_0$  and  $\bar{g}_0$  defines the *deflection of the vertical*  $\theta$ . In other words, the deflection of the vertical is the angle between the actual plumb line taken on the geoid and the ellipsoidal normal, and is usually referred to as the *absolute deflection* on the geoid. Analogous to geoidal height, a *relative deflection* can be defined (see §21.1) by using an ellipsoid that is not geocentric.

Some applications require that deflections be defined on the surface of the earth. In these instances, the angle of interest is between the actual gravity vector taken on the earth's surface and the ellipsoidal normal: it is called the *surface deflection*. Again, it can be either absolute or relative according to which kind of ellipsoid is

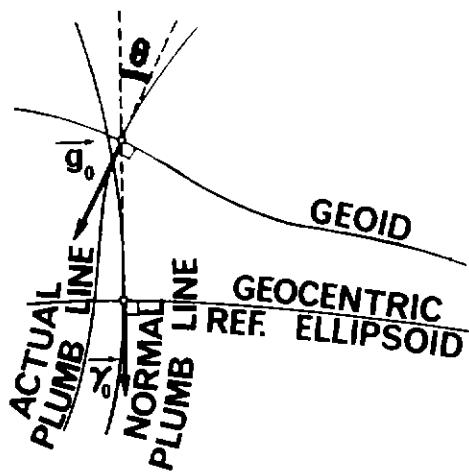


FIG. 6.19. Deflection of the vertical.

used (geocentric or non-geocentric). Evidently, surface and geoid deflections for the same point (the same ellipsoidal normal) are different, because the actual plumb lines between the geoid and the earth's surface are curved and twisted. As a rule, the differences are expected to be more significant in mountainous regions. In the Alps, values of up to  $12''$  have been computed and reported by KOBOLD AND HUNZIKER [1962]. More about this will be found in §21.3.

Since the deflection of the vertical  $\theta$  is a spatial angle, it is convenient to decompose it into two orthogonal components  $\xi$  and  $\eta$ , called the *deflection components*. FIG. 20 shows one-eighth of a sphere centred at the point ( $T$ ) for which the deflection is given (either on the earth's surface or on the geoid). The Cartesian

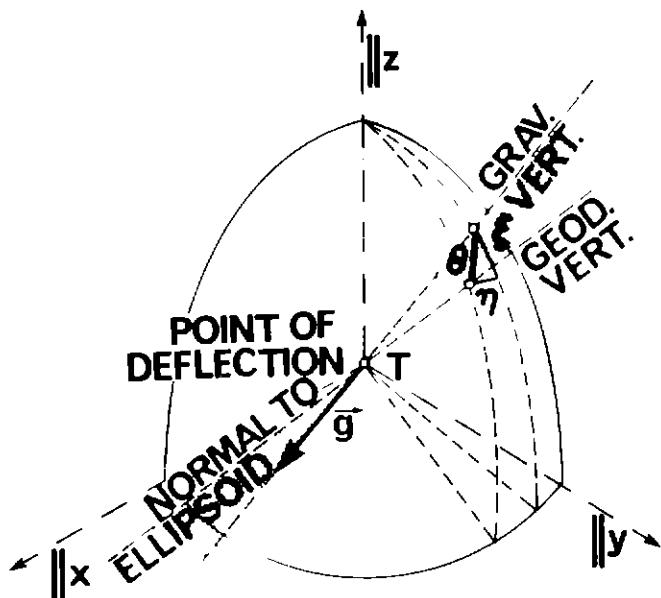


FIG. 6.20. Deflection components.

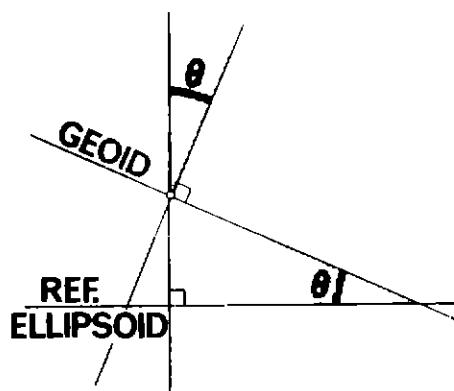


FIG. 6.21. Deflection as the geoidal slope.

coordinate system shown is parallel with the natural system of coordinates introduced in §5.3. Thus  $\xi$  is the projection of  $\theta$  onto the meridian plane, and as such is often called the *meridian component*;  $\eta$  is the projection onto the prime vertical plane, and as such is called the *prime vertical component*.

The sign of both components is conventionally adopted as positive, if the actual gravity vertical is farther north and farther east than the geodetic vertical. This is the general rule for both the Northern and Southern Hemispheres. It is not difficult to realize that the geoidal deflection  $\theta$  can be viewed as the maximum slope of the geoid

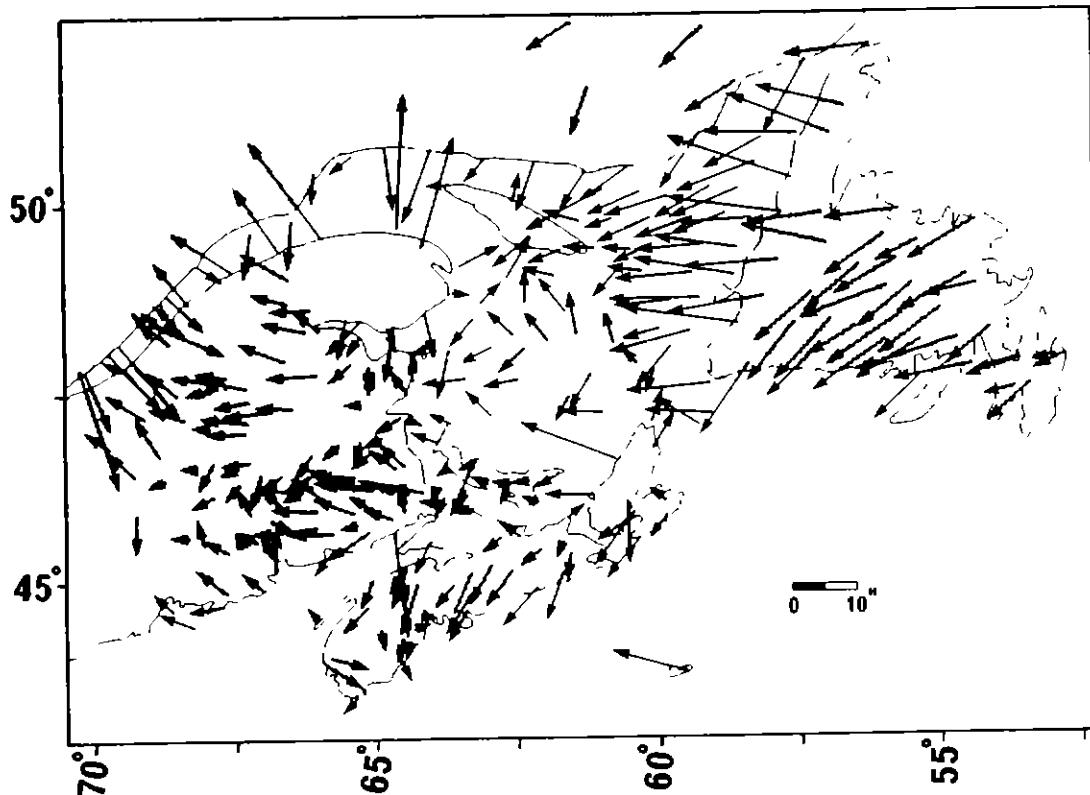


FIG. 6.22. Local relative deflections for Eastern Canada referred to the NAD 27.

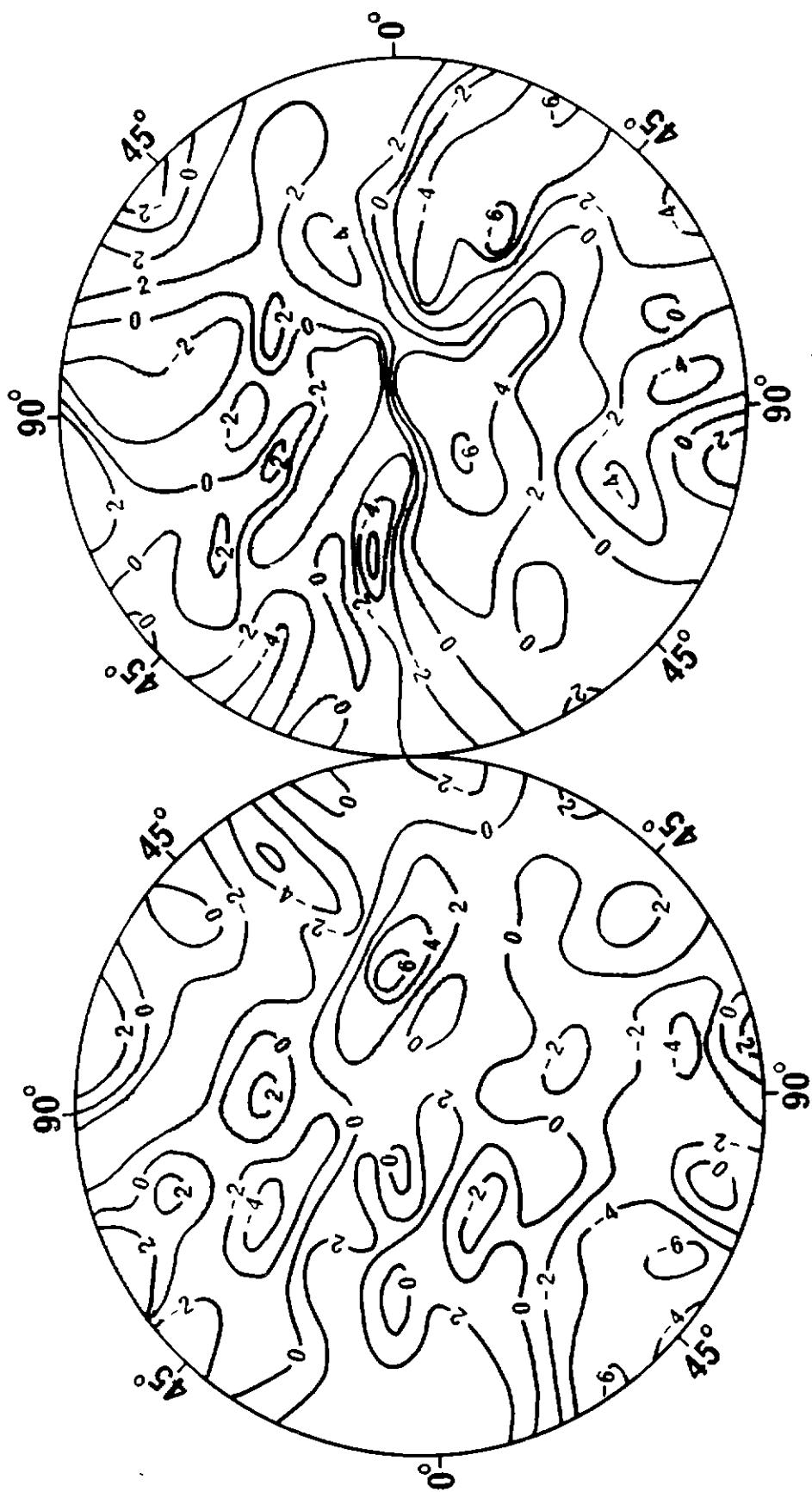


FIG. 6.23. Global absolute meridian deflection components. Contours in seconds of arc.

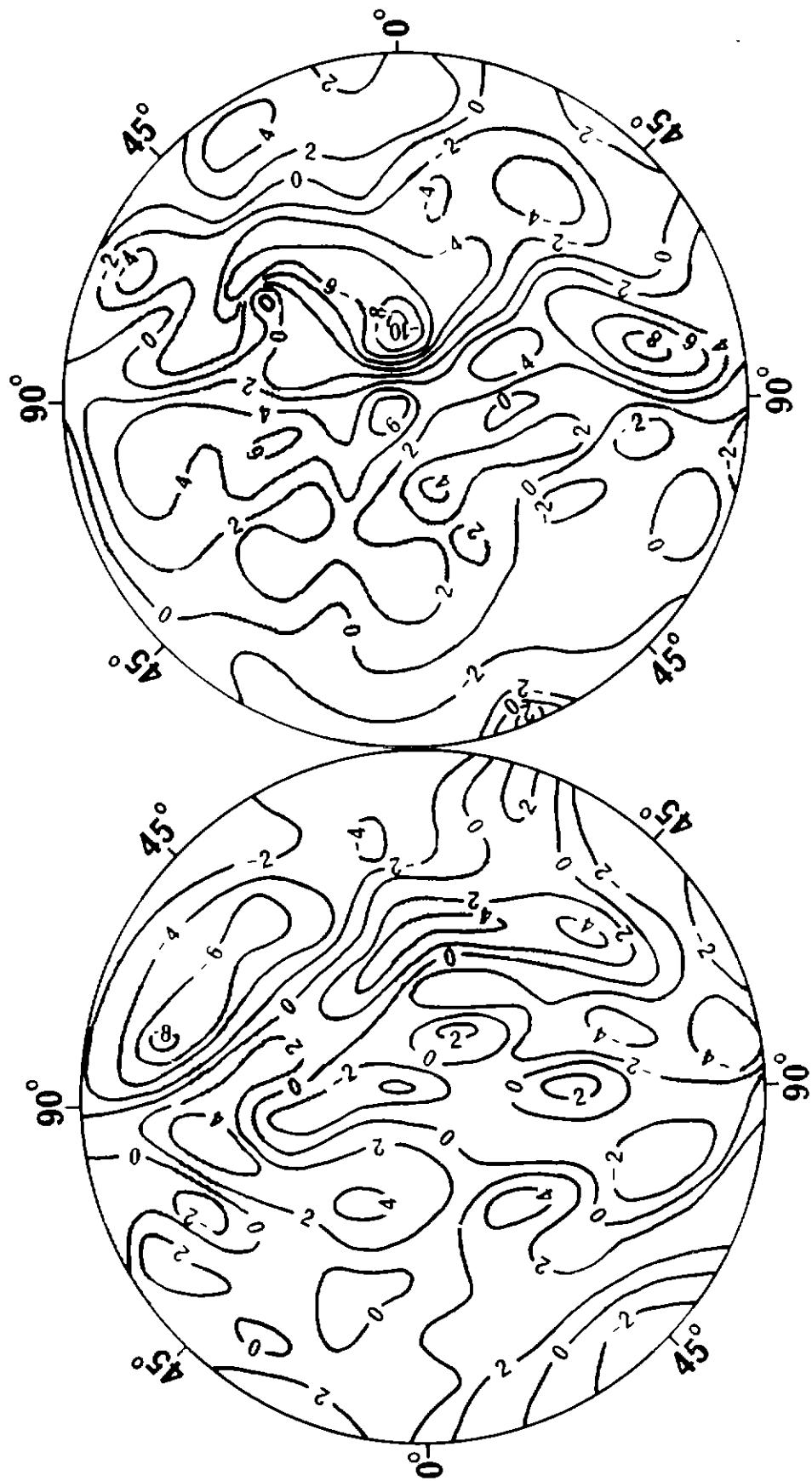


FIG. 6.24. Global absolute prime-vertical deflection components. Contours in seconds of arc.

with respect to the reference ellipsoid (cf. FIG. 21). Similarly,  $\xi$  is the geoidal slope in the north-south direction and  $\eta$  is the slope in the east-west direction. If the geoid slopes down towards north, then  $\xi$  is positive and vice versa. A downward geoidal slope east corresponds to  $\eta$  positive. In North America, the sense of  $\eta$  is sometimes reversed; this practice will not be followed here.

The deflections can be regarded as being composed of two parts: one reflecting the regional density distribution, and the other reflecting the complexity of the earth's surface topography and close-by density anomalies. The first part is generally predominant in flat and low areas; the second in mountainous reliefs. The largest surface deflections are found in high mountains, where values of the order of  $1'$  may be encountered [HEISKANEN AND VENING MEINESZ, 1958]. Deflections are seldom larger than  $20''$  in the lowlands.

Local, relative deflections for Eastern Canada (related to a non-geocentric ellipsoid) are shown in FIG. 22 [MERRY AND VANÍČEK, 1974]. Smoothed global (absolute) deflection components, related to the IAG 1967 reference ellipsoid, are given in FIGS. 23 and 24. These were obtained from observations to satellites in the 1960s [BURŠA, 1971]. The reader may want to compare the magnitudes of the two kinds of deflections. The much smaller magnitudes of the global deflections are the consequence of the smoothing of the global field. Such was also the case with global gravity anomalies and the global geoid.

## CHAPTER 7

### EARTH AND ITS SIZE AND SHAPE

The determination of the size and shape of the earth is one of the main tasks of geodesy. Hence, a clear understanding of the different meanings of the terms size and shape, as they pertain to geodesy, is essential. In geodesy, when speaking about the earth's figure, the earth is normally regarded as being rigid. The time perturbations of the size and the shape are then treated separately. This is the approach we use here.

Each of the following four sections will describe one of the four different kinds of surfaces used in geodesy to model the geometry of the earth's surface. The first section is concerned with the most natural surface which is the physical surface of the earth—the terrain as well as the surface or the bottom of the oceans. This surface is complicated and, as such, is inconvenient to work with mathematically. The next surface that has a definite physical interpretation is the geoid, already introduced in the previous chapter and discussed here in the second section. The geoid is a much smoother surface than the terrain, but it is still too complicated to serve as a useful computational surface on which to solve geometrical problems like positioning. More convenient, from that point of view, are ellipsoidal surfaces, which are discussed in the third section. The last section is devoted to other, more complex, mathematical surfaces used for solving other geodetic problems.

#### 7.1. Actual shape of the earth

When portraying the shape of the physical surface of the whole or a part of the earth, use is made of graphical or digital topographical maps of various types and scales. This is the interest that geodesy and mapping have in common. At present, about 72% of the earth's surface is covered with water, and only the remaining 28% is dry land [GASS ET AL., 1972]. For obvious reasons, our interest is focussed on this 28 percent. Although it is primarily the geometry of the terrain that geodesy aims at describing, the sea surface and bottom are by no means excluded.

To describe the terrain mathematically, one may choose a finite set of points representative of the terrain, monument them, and give their positions in a selected coordinate system. Networks of these points may then be thought of as one possible representation of the earth's physical surface. Traditionally, these *geodetic networks* fall into three categories, depending on how the positions of individual points are

defined. Networks of points defined by only one coordinate, the 'height above sea level'  $H$ , are known as geodetic height networks, sometimes also called 'vertical networks'. Networks of points with known horizontal positions, say latitude  $\phi$  and longitude  $\lambda$ , are called geodetic horizontal networks. The reason for this split into horizontal and height networks is mostly historical. In the past, it was easier and more economical to determine horizontal and vertical positions separately. They each require different kinds of field observations; also, they only affect each other weakly (cf. Part IV). Finally, networks of points positioned by three coordinates are referred to as three-dimensional networks. Let us now enlarge upon these statements.

It is evident that even the height network points, usually called *bench marks*, must have some horizontal positions associated with them so that their location on the earth is known at least approximately. The main difference between the height and horizontal network points lies in the fact that for height points the horizontal position is known only weakly, and for horizontal points the vertical position is determined only approximately. Thus the emphasis, as far as accuracy is concerned, is placed on the vertical positions in height networks and on the horizontal positions in horizontal networks.

*Geodetic height networks* are divided into networks of different orders. The higher the order, the higher the accuracy and, generally, the larger the spacing between adjacent bench marks. The spacing varies from country to country with the average for the first (highest) order being about 1 to 2 kilometres. The spacing between adjacent lines of bench marks, however, is usually greater. A global view of first-order networks established before 1970 is given in FIG. 1 [U.S. ARMY TOPOGRAPHIC COMMAND, 1971a].

Vertical positions (heights)  $H$  are commonly given with respect to the sea level or, more precisely, with respect to the geoid called, in this context, the *vertical geodetic*

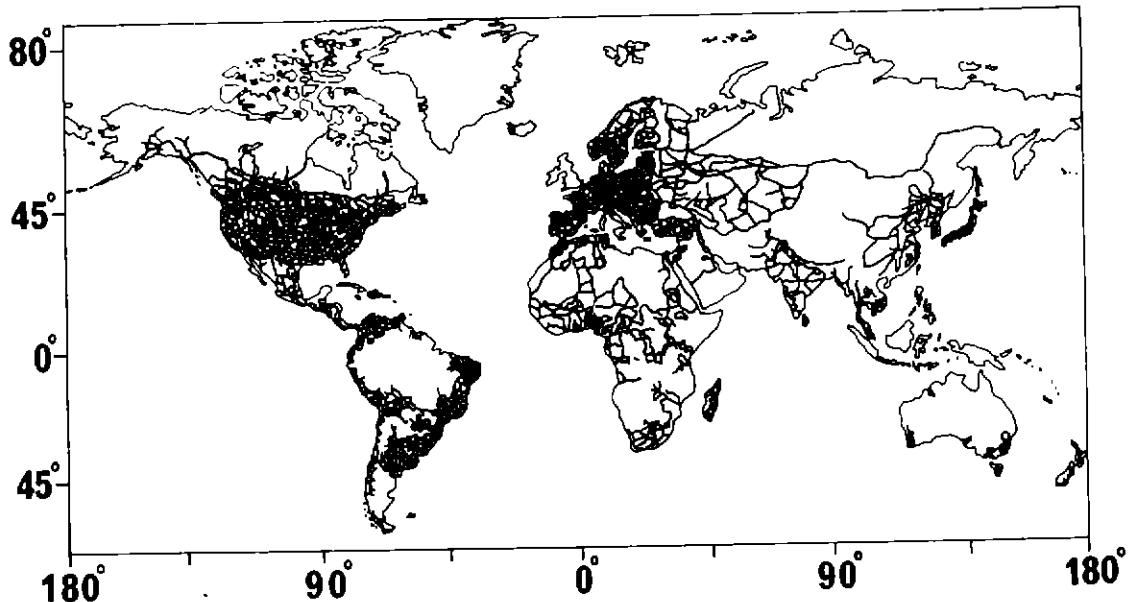


FIG. 7.1. Geodetic height networks.

*datum* It can then be said that heights are referred to the geoid or, more popularly, we speak about heights above the sea level. The land measurements are connected with the sea level through the observations of sea level positions using instruments known as tide gauges (for details see §19.1). Other reference surfaces may also be used, as will be seen in §16.4.

The heights of millions of first-order height control points have been determined all over the surface of the earth. These heights are known to an accuracy better than a metre, while the accuracy of height differences of adjacent bench marks is much higher. According to NASA [1973], the standard deviation  $\sigma$  in the height  $H$  above the geoid propagates with distance  $S$  and obeys the following global empirical formula:

$$\sigma_H = 1.8 \times 10^{-3} S^{2/3} \text{ metres,} \quad (7.1)$$

where  $S$  is in kilometres. Thus the difference between the heights of two points 2000 km apart is accurate to  $\sigma_H \approx 0.3$  metres. At present, the height is the most precisely known geodetic coordinate.

Geodetic horizontal networks consist of monumented points the geodetic ellipsoidal coordinates  $\phi$  and  $\lambda$  (cf. §3.3) of which are known. They are referred to a reference ellipsoid called the *horizontal geodetic datum*. These horizontal positions may be given in any other two-dimensional coordinate system, such as a mapping coordinate system, whose relation to the reference ellipsoid is known. In geodetic practice, conformal mappings are normally used (see §16.3). The heights of these horizontal control points are determined only approximately or not at all.

Horizontal networks are divided into networks of different orders according to the accuracy of the horizontal positions. First-order networks have the highest accuracy of 'relative' positions (i.e., of horizontal coordinate differences of adjacent points) of about  $1/100000$ , i.e.,  $\sigma = 10$  cm in a coordinate difference of 10 kilometres. Some modern, rigorously established networks, however, may have an accuracy several times higher. Lower orders have lower accuracies, and these vary from country to country. The order is also usually reflected in the spacing of adjacent points. Higher order networks are composed of points further apart, while the density increases in lower order networks. Typically, limiting ourselves to terrestrial techniques alone, spacing in the first-order networks is in the tens of kilometres with the station intervisibility and clearance above the ground being the main limiting factors.

Similar to the case of height networks, horizontal networks are an example of differential measurements used in an integration process. Consequently, observing errors accumulate in spite of all possible precautions. Inevitably the accuracy of the 'absolute' positions is worse than that of the relative positions and deteriorates with distance from a preselected origin of the network (cf. §18.1). For instance, for the horizontal networks in the United States, SIMMONS [1950] found that the standard deviation  $\sigma_p$  of the difference of coordinates between a point in the network and the origin (Meade's Ranch, Kansas) is given by the following empirical formula:

$$\sigma_p = 4 \times 10^{-2} S^{2/3} \text{ metres,} \quad (7.2)$$

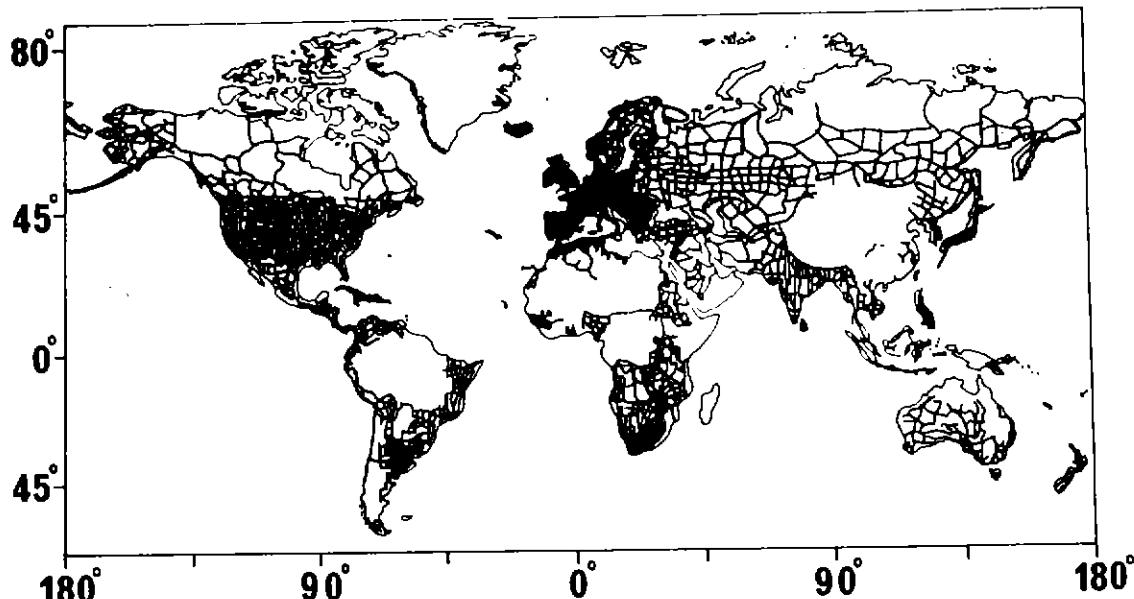


FIG. 7.2. Geodetic horizontal networks.

where  $S$ , the distance between the two points, is in kilometres. This indicates, e.g., that a point located 2000 km from Meade's Ranch has an error  $\sigma_p = 6.4$  m associated with its position taken with respect to Meade's Ranch. The numerical factor in (2) is specific to the networks in the United States. In other parts of the world, a different factor would apply because of different observing and computing techniques.

FIG. 2 shows the global view of geodetic horizontal networks established from terrestrial observations prior to 1970 (according to the U.S. ARMY TOPOGRAPHIC COMMAND [1971b]). It is worth noting that these networks have large gaps. Further, the networks are referred to scores of different reference ellipsoids, whose mutual positions are still largely unknown. Thus the information the networks give about the earth's shape and size is of limited value. More about horizontal geodetic datums will be found in §7.3. Here only regions using the main geodetic datums, according to NASA [1973], are shown (FIG. 3).

After becoming acquainted with the height or horizontal networks, one realizes that their value is impaired because they portray only one or two dimensions of the inherently three-dimensional figure of the earth. In addition, the overlap of height and horizontal control points is, unfortunately, minimal. Points conveniently placed for horizontal positioning (usually the tops of hills) are unsuitable for vertical positioning (points normally located along roads and railways) and vice versa.

From the geodetic standpoint, it is natural to work with a network of points whose three-dimensional coordinates are known. Such *three-dimensional networks* can be established using either one of the following two approaches:

(a) Combine the known horizontal  $(\phi, \lambda)$  and vertical ( $H$ ) positions of corresponding points to obtain the triplet of coordinates  $(\phi, \lambda, h)$  or, equivalently (see §3.3), the triplet  $(x, y, z)^G$ . For a full explanation of this coordinate system see §15.4. Clearly, the knowledge of geoidal height  $N$  is indispensable to the evaluation

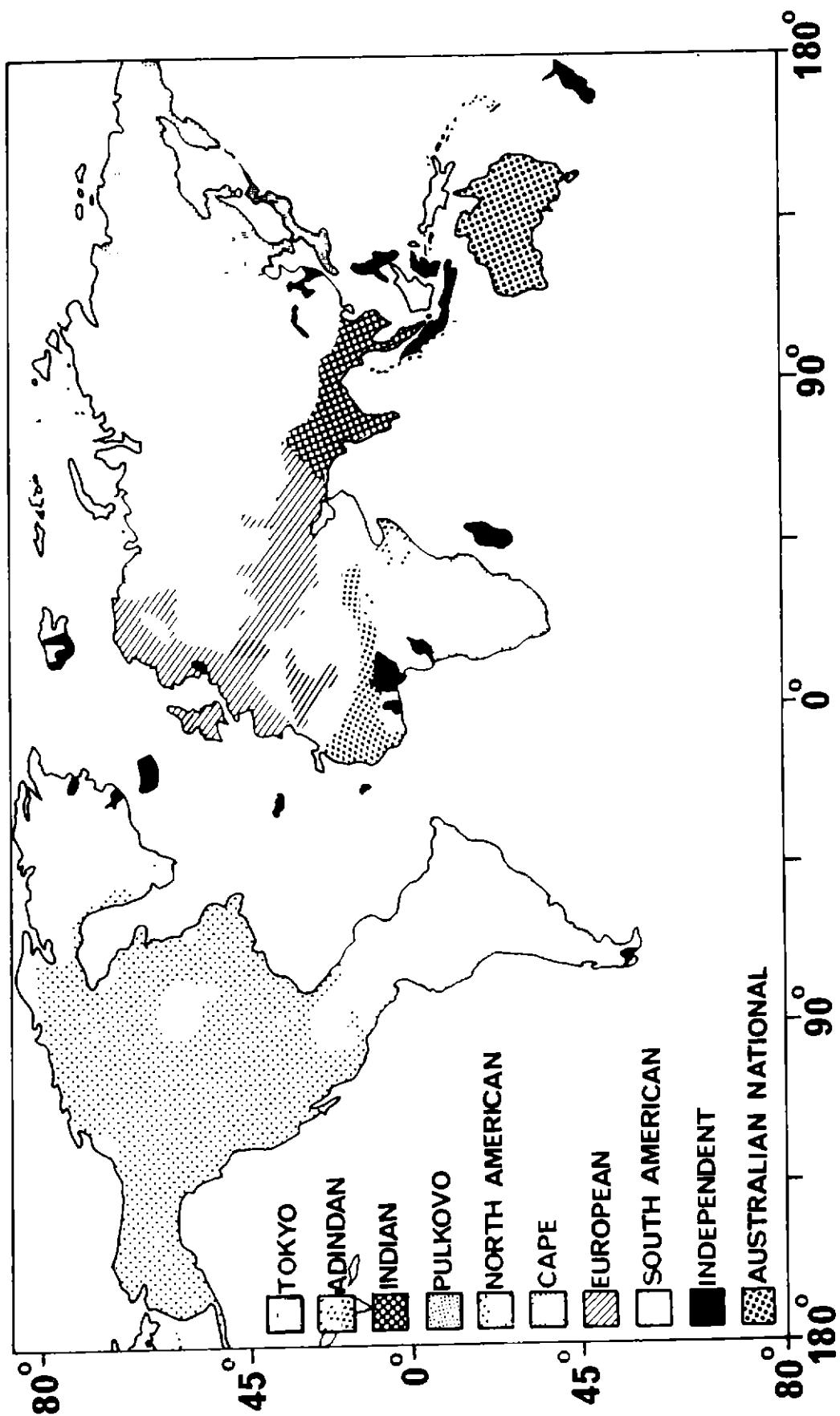


FIG. 7.3. Regions of use of major geodetic datums.

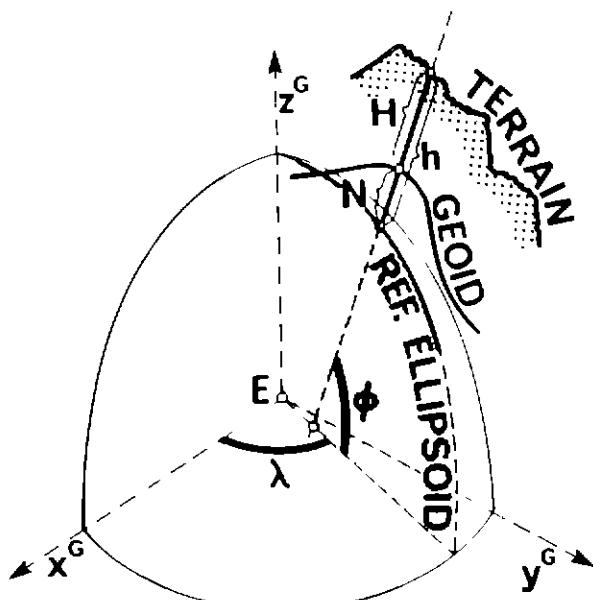


FIG. 7.4. Geoid as the link between one-, two-, and three-dimensional positions.

of the height  $h$  above the ellipsoid (cf. FIG. 4):

$$h = H + N . \quad (7.3)$$

It has only been in the past decade or so that sufficient data have permitted an accurate enough determination of geoidal height for this purpose.

(b) Use terrestrial or extraterrestrial positioning techniques that give three-dimensional positions directly. One such terrestrial technique is the high precision geodetic traversing originally conceived of as calibration for satellite positioning. Analytical photogrammetry and inertial positioning are illustrations of terrestrial positioning techniques suitable for the densification of three-dimensional networks. In this context, satellite positioning appears to be particularly versatile. One of its main advantages is that the interstation distances can be increased arbitrarily since intervisibility is no longer a requirement. All these techniques will be treated in Part IV.

One of the first systematic attempts to use satellite technology—namely, photographs of satellites against the star background—for establishing a global, geodetic, three-dimensional network [SCHMID, 1974] resulted in the acquisition of coordinates for the points shown in FIG. 5. The accuracy of the absolute positions of these points is said to be of the order of or better than  $\sigma = 5$  m for all three coordinates. Relative positions of points, using more up-to-date satellite positioning technology, can now be determined with a submetre accuracy, and prospects are good for further improvement by at least one order of magnitude. An example of a regional, three-dimensional network established by employing a technique based on the Doppler effect is given in FIG. 6 [KOUBA AND BOAL, 1976].

This concludes our brief outline of dry land geodetic networks that are used as a tool for determining the earth's shape. Traditionally, geodetic networks had been

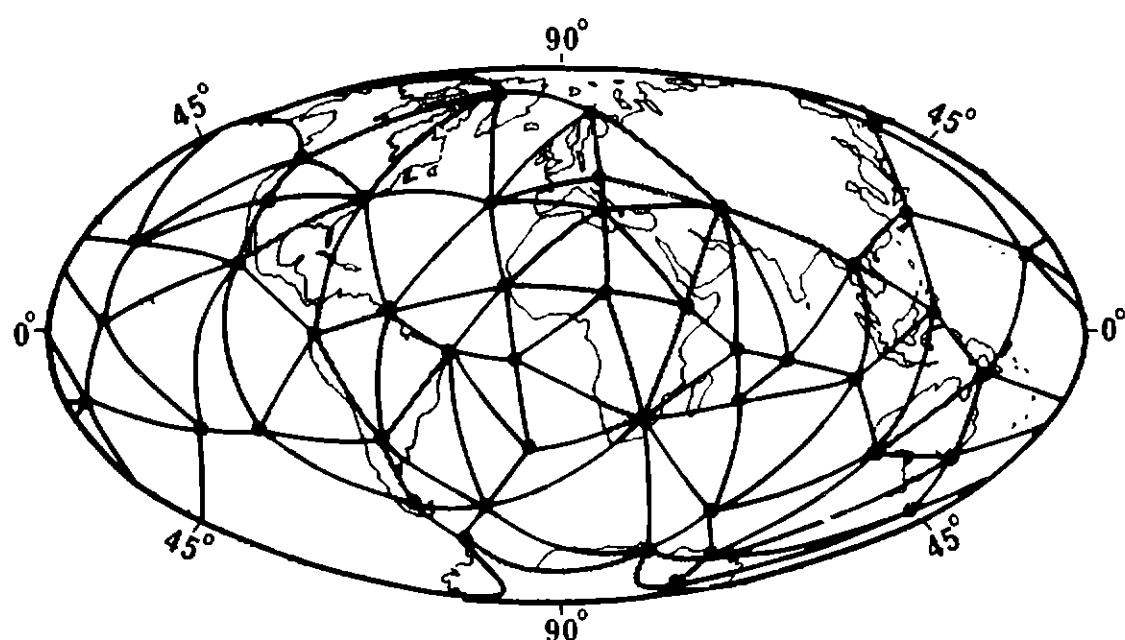


FIG. 7.5. BC-4 worldwide satellite triangulation network.

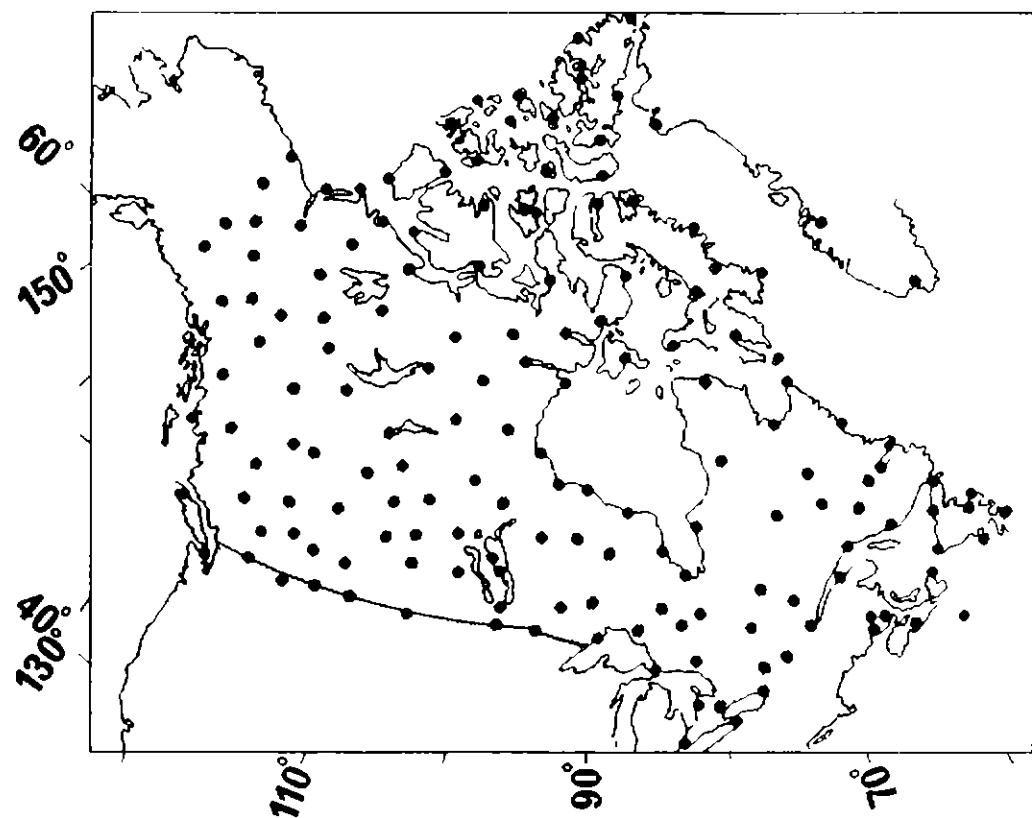


FIG. 7.6. Canadian Doppler satellite network.

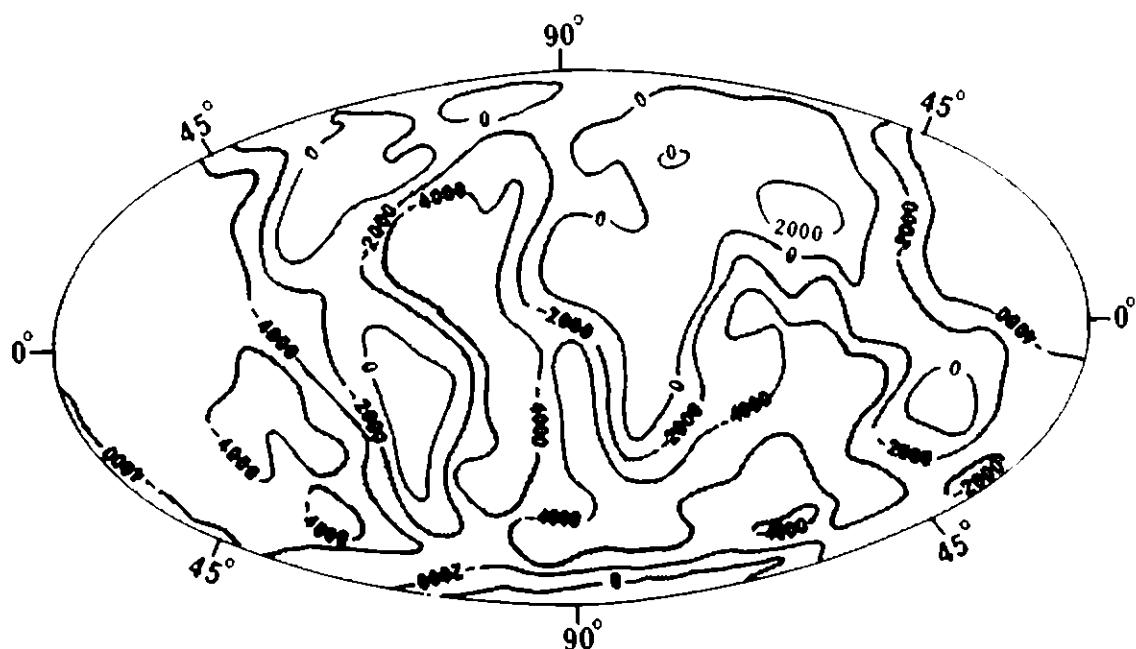


FIG. 7.7. The earth's topography in spherical harmonics. Contours in metres.

confined to dry land, due to the lack of technology for offshore positioning as well as the lack of motivation, because the exploitation of the sea floor has been very limited. Recently, the technology for these tasks has undergone a rapid development, and it is now both desirable and feasible to extend geodetic networks to the sea floor [MCKEOWN AND EATON, 1974; CESTONE ET AL., 1976]. On the other hand, since bathymetric charts have always been needed for navigation, hydrographical surveying has been practiced, in one form or another, since the early Greeks [HYDROGRAPHER OF THE NAVY, 1965]. The charting of sea depth is routinely carried out by various hydrographical institutions around the world.

For some purposes, the earth's shape has to be expressed through a mathematical formula. If so, a *continuous representation of the earth's surface* is obtained, as opposed to the discrete, point representation treated above. Evidently, because the earth's physical surface is very rugged at places, one would have to use very long functional series to get a reasonably accurate expression. Attempts have been made to portray at least the more prominent features using functional series of sensible length. FIG. 7 shows the results of one such representation [PREY, 1922]. Note that in this figure, the bottom of the seas, rather than the sea level, is shown. A more detailed portrait can be found in LEE AND KAULA [1967].

## 7.2. Geoid as a figure of the earth

In agreement with Gauss's idea, the geoid is often regarded as another representation of the figure of the earth. As seen in §6.4, the geoid is a surface with a definite physical meaning; being an equipotential surface, the geoid describes the surface of homogeneous water. With sea water being more or less homogeneous, the geoid ap-

proximately follows the sea level and, thus, closely represents the figure of the earth in 72% of the terrestrial globe.

Let us now clarify what is meant by the 'approximate coincidence' of the geoid and the sea level. Of course, one realizes that the sea surface is much less stable than the land surface. It changes continuously with time under the influence of various effects (such effects will be discussed in detail in §8.4 and §19.1). Is it then plausible to talk about a coincidence of the geoid and the sea level? To be able to compare these two surfaces in a stationary manner, as is done with the geoid and the terrain, a stationary surface called the *mean sea level* is used.

Observations of the instantaneous sea level show that, while it may vary by as much as a couple of tens of metres within one day, the monthly averages do not vary more than several decimetres, and annual averages are usually stable to some 10 cm within a period of several decades [HILL, 1966]. This fact gave rise to the idea of defining the mean sea level (MSL) as the average of all the instantaneous sea surfaces over a long period of time. The shape of the mean sea level is thus determined from records of the instantaneous sea level at a chain of observational points located on coastlines and equipped with tide gauges. The Permanent Service for Mean Sea Level (PSMSL), with headquarters at Bidston Observatory (Cheshire, U.K.), has been charged with the collection and dissemination of worldwide mean sea level data [PSMSL, 1976].

The global sea water distribution displays secular (semipermanent) inhomogeneity due to the shape of coastlines and other factors, such as the permanently higher temperature of waters in the equatorial belt and lower temperature in the polar regions, the prevailing pattern of geostrophic winds, etc. Many of these effects can be measured, and appropriate departures of the mean sea level from a hypothetical state corresponding to an undisturbed homogeneous fluid can be at least approximately estimated. These departures are often called the *sea surface topography* to emphasize the analogy with dry land topography, i.e., the departure of the land surface from the geoid. Several attempts to determine the sea surface topography have been made. FIG. 8 shows the results obtained by HELA AND LISITZIN [1967]. This figure should also help us in understanding the proper meaning of the definition of the geoid (cf. §6.4) as the equipotential surface passing, in the mean sense, through the (stationary) mean sea level.

Let us now return to the relation between the geoid and a geocentric reference ellipsoid. It can be seen from FIG. 6.17 that, relatively speaking, the geoid follows the geocentric reference ellipsoid very closely. The maximum geoidal height is of the order of 100 m which, compared with the mean radius of the earth ( $R \doteq 6371$  km), represents about  $1.6 \times 10^{-5} R$ . Thus in many geodetic applications, where geoidal height  $N$  comes into various corrective terms also containing  $R$ ,  $N$  is treated as a relatively small quantity. If, on the other hand, a best-fitting reference sphere were used, geoidal heights referred to it would be as large as 10.7 km, i.e.,  $1.7 \times 10^{-3} R$ , or more than a hundred times larger than when the biaxial ellipsoid is used. Thus the approximation of the geoid by the geocentric reference ellipsoid is two orders of magnitude better than that of a sphere.

Similar to the terrain, the geoidal surface can be portrayed either in a discrete fashion, as a set of points, or in a continuous mode, in terms of a mathematical

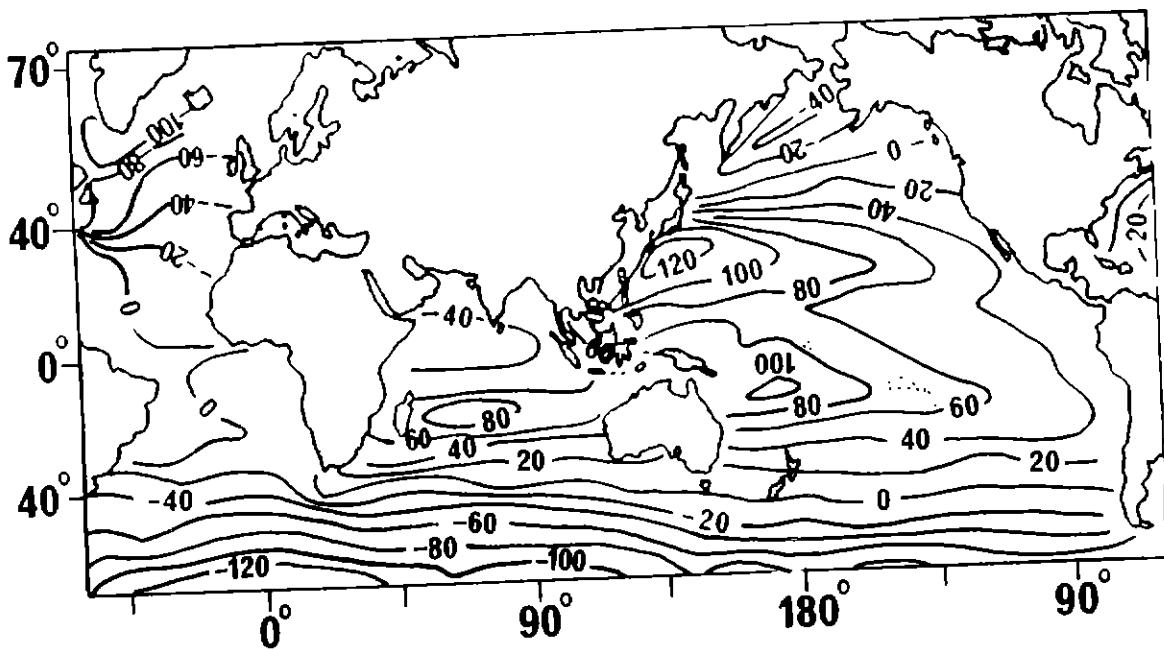


FIG. 7.8. Sea surface topography. Contours in centimetres.

formula. Even though the geoid is convex everywhere, it is still a very complicated surface, and only an infinite functional series can describe it exactly. In practice, finite (truncated) series are used to represent the geoid approximately. One such example was seen in FIG. 6.17; more will be said about this point in Chapter 20.

As will also be seen later (Part V), the shape of the geoid, relative to a reference ellipsoid, can now be determined accurately to about 1 to 2 m in regions, such as Europe, North America, and Australia, where adequate data coverage exists. Geoidal height differences at points not more than a few hundred kilometres apart can be obtained more accurately. The accuracy of geoidal heights in other parts of the world is lower, the lowest being in the polar regions, because of the lack of data.

For the two-dimensional, geometrical computations necessary in horizontal positioning, it is convenient to use a surface that is close to the geoid but still as simple as possible to mathematically describe. It should be evident from the explanations that the geoid itself is far too complex a surface to be useful for these computations and, as such, is not convenient as a reference surface for geodetic horizontal networks. It was proposed by some researchers to use a simplified version of the geoid for this purpose. The various simplifications would be expressed by just a few first major terms taken from the functional series describing the geoid. These simplified surfaces became known as *spheroids*. Two such spheroids were worked out, one by Helmert and the other by Bruns [HEISKANEN AND MORITZ, 1967], but neither one found wide usage. While being geometrically more complex than a biaxial ellipsoid, they do not deviate from a biaxial reference ellipsoid by more than some tens of metres. In fact the triaxial and biaxial ellipsoids, to be discussed next, can be regarded as spheroids of a lower order.<sup>1</sup>

<sup>1</sup>It is because of the necessity to distinguish these surfaces that, in geodesy, the term ellipsoid is used for the simple 2 or 3 parametric surfaces of the second order, which many other sciences customarily call spheroids.

A mathematically viable surface, closest to the geoid, is a *triaxial ellipsoid*. Many researchers have estimated the parameters of a triaxial ellipsoid that would best approximate the geoid. Such a triaxial ellipsoid has three mutually perpendicular axes positioned in the earth as follows: the minor, coinciding with the earth's principal (polar) axis of inertia; the major, and the medium, both lying in the equatorial plane. Thus, the triaxial ellipsoid is defined by the lengths of the major axis ( $2a$ ), the minor axis ( $2b$ ), the medium axis ( $2c$ ), and the orientation of the major axis in the equatorial plane.

Usually, the following are taken as the four defining parameters:

- (a) Half length of the major axis  $a$ .
- (b) Polar flattening  $f$ , given by

$$f = \frac{a - b}{a} \quad (7.4)$$

- (c) Equatorial flattening  $f_e$ , given by

$$f_e = \frac{a - c}{a} \quad (7.5)$$

- (d) Geographical longitude  $\lambda_a$  of the major axis.

Three of the better known determinations, using different kinds of data, are given in TABLE 1.

As an illustration, the fit of Burša's ellipsoid to the geoid is given in FIG. 9. The most noticeable feature is that this geoid does not look much smoother than one referred to the biaxial ellipsoid (cf. FIG. 6.17). This is explained by the fact that the equatorial flattening is very small compared with the polar flattening. (Clarke's and Heiskanen's solutions show unrealistically large values of equatorial flattening.) This means that, like the spheroids, the best-fitting triaxial ellipsoid deviates only by a few tens of metres from the corresponding biaxial ellipsoid, whose two equatorial axes are equal. The large spread in the solutions for the direction of the major axis illustrates this point.

As another example, the actual shape of the geoid in the equatorial and one of the meridian cross sections is given in FIGS. 10 and 11. These cross sections, taken from FIG. 6.17, show that the departures of the geoid from the geocentric (biaxial)

TABLE 7.1  
Triaxial ellipsoids

Solution	$a$ (km)	$f^{-1}$	$f_e^{-1}$	$a - c$ (m)	$\lambda_a$ (degrees)
CLARKE [1878]	6378.206 <sup>a</sup>	293.2	13 720	465	8 west
HEISKANEN [1938]	6378.388 <sup>a</sup>	297.8	18 120	352	23 west
BURŠA [1971]	6378.173	297.78	92 800	68	14.8 west

<sup>a</sup>Value not explicitly stated in the original publication.

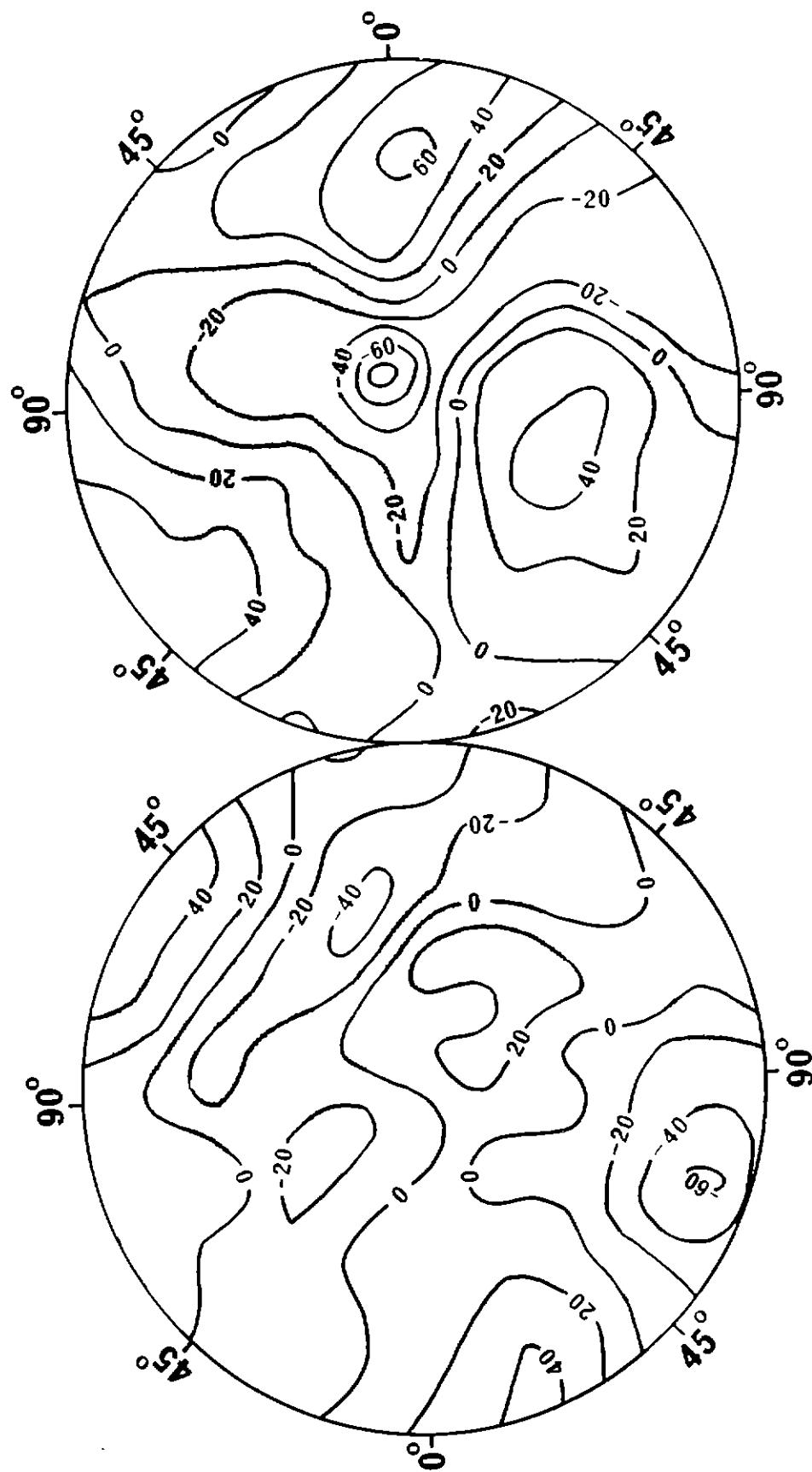


FIG. 7.9. Global geoid as referred to a triaxial ellipsoid. Contours in metres.

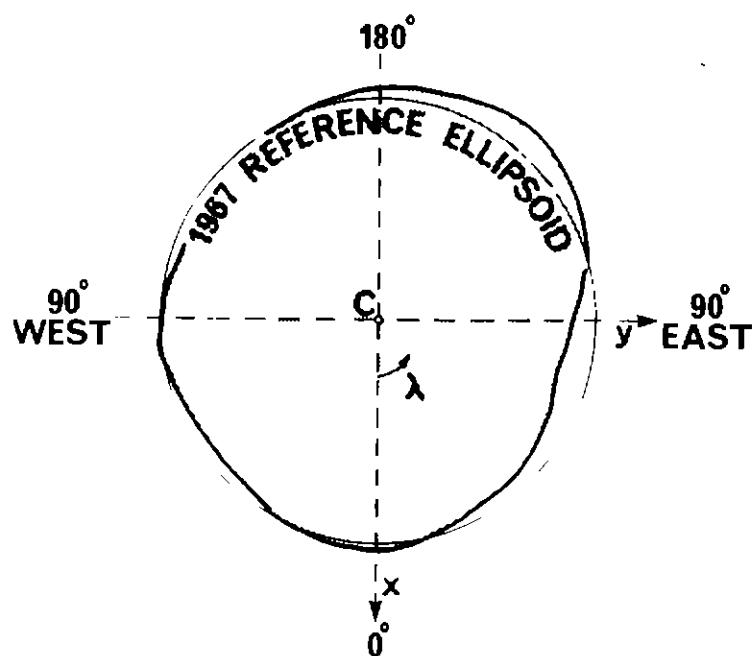


FIG. 7.10. Equatorial cross section of the geoid. (Scale of departures from the ellipsoid exaggerated  $10^4$  times.)

reference ellipsoid in the equatorial plane are not significantly different from those in the meridian planes. While the departures in the equatorial plane would diminish if a triaxial ellipsoid was used, similarly the departures in the meridian planes would diminish if a *pear-shaped reference body* was taken instead of the biaxial ellipsoid. Hence, one can argue that a pear-shaped reference body might be as appropriate as a triaxial ellipsoid.

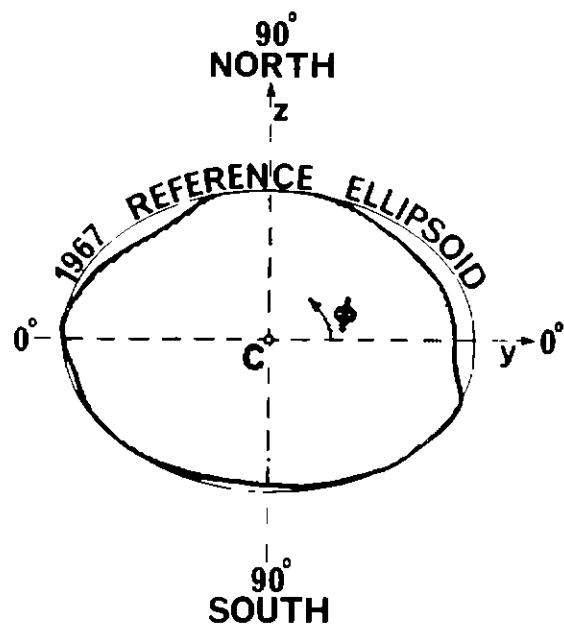


FIG. 7.11. Meridian cross section of the geoid along  $\lambda = 90^\circ$  meridian. (Scale of departures from the ellipsoid exaggerated  $10^4$  times.)

### 7.3. Biaxial ellipsoid as a figure of the earth

As we have already seen, neither the best-fitting triaxial ellipsoid nor the best-fitting pear-shaped body departs much from the *biaxial ellipsoid*. But, the computations on a triaxial ellipsoid are much more complicated than those on a biaxial ellipsoid. Therefore, in practice, biaxial ellipsoids are used exclusively with the understanding that, although they do not fit the geoid as well as the triaxial ellipsoids do, this disadvantage is more than offset by the ease they provide for computations. Biaxial ellipsoids are uniquely defined by only two parameters—usually the major semi-axis  $a$  and flattening  $f$ —given by (4).

The problem of finding the best-fitting biaxial ellipsoid is the classical geodetic problem that has intrigued scientists for centuries. Let us outline here, at least superficially, the main techniques used in the past to solve this problem. Eratosthenes (see §1.1) assumed the earth to be spherical, i.e.,  $f = 0$ , and derived the radius of the earth, thought now to have been about 5950 km, i.e.,  $a \doteq 5950$  km [SCHWARZ, 1975], from a direct measurement of the length of a meridian arc and the corresponding latitude difference.

Based on a similar principle was another famous determination commissioned by the French Academy of Sciences (see §1.2). It was felt that, by showing the concepts of the French experiment here, some light would be shed on the eighteenth century thinking, because repercussions of these concepts influenced the geodetic way of thinking for well over a century. In order to determine the two parameters,  $a$  and  $f$ , two meridian arcs—one in Peru and the other on the border between Sweden and Finland—were observed. The length  $S_{ij}$  of each arc was derived from a relatively

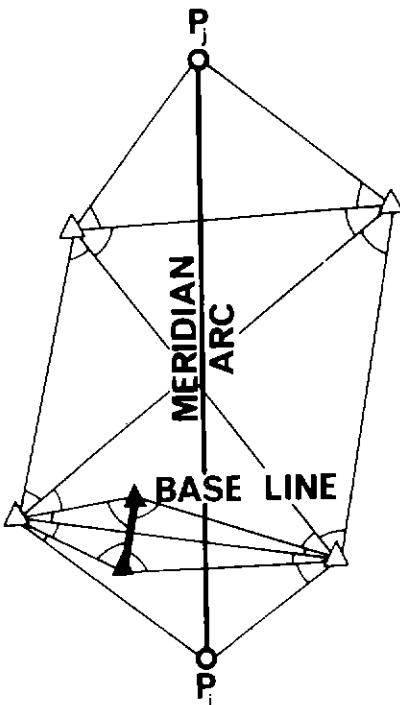


FIG. 7.12. Determination of a meridian arc length.

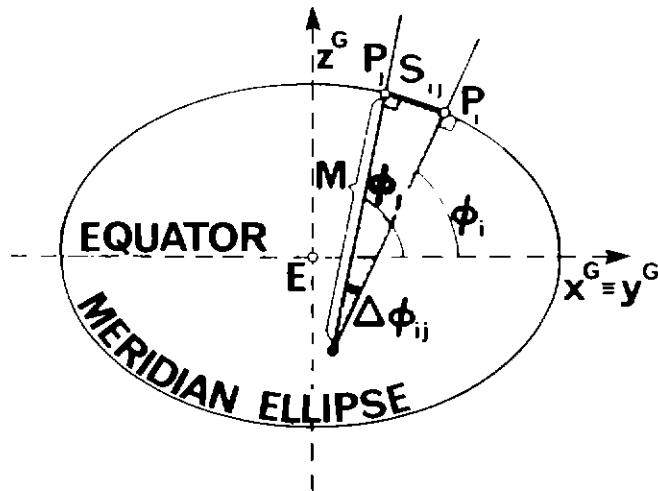


FIG. 7.13. Geometry of a meridian ellipse. (For the meaning of superscript G, see §15.4.)

short, directly measured base line through a configuration of triangles where all horizontal angles were measured (triangulation). One such typical configuration is shown on FIG. 12. The latitudes of the end points  $P_i$  and  $P_j$  were determined from astronomical observations. Once the latitudes  $\phi_i$ ,  $\phi_j$  and the corresponding length of the meridian arc  $S_{ij}$  are known, the pertinent equation can be formulated, using the geometry of the meridian ellipse that shares the major semi-axis  $a$  and flattening  $f$  with the sought biaxial ellipsoid (see FIG. 13).

The meridian arc length  $S_{ij}$  is related to the latitudes  $\phi_i$  and  $\phi_j$  by

$$S_{ij} = \int_{\phi_i}^{\phi_j} M(\phi) d\phi, \quad (7.6)$$

where  $M(\phi)$  is the radius of curvature of the meridian ellipse, or simply *meridian radius of curvature*, that changes with latitude  $\phi$ . From differential geometry, it is known (cf. §3.3) that  $M(\phi)$  is given by the following formula:

$$M(\phi) = \frac{dS}{d\phi}. \quad (7.7)$$

To develop the equation for  $M(\phi)$  involving the defining parameters  $(a, f)$  of the ellipsoid, let us first derive the equation for the  $p$  coordinate of point  $P_i$ , i.e.,  $p_i^G$  (see FIG. 14). The equation of an ellipse, in Cartesian coordinates  $p^G$ ,  $z^G$ , is

$$\frac{(p^G)^2}{a^2} + \frac{(z^G)^2}{b^2} = 1. \quad (7.8)$$

On the other hand, the relation between the latitude and the Cartesian coordinates of a point on the ellipse is given by

$$\tan \phi = - \frac{dp^G}{dz^G}. \quad (7.9)$$

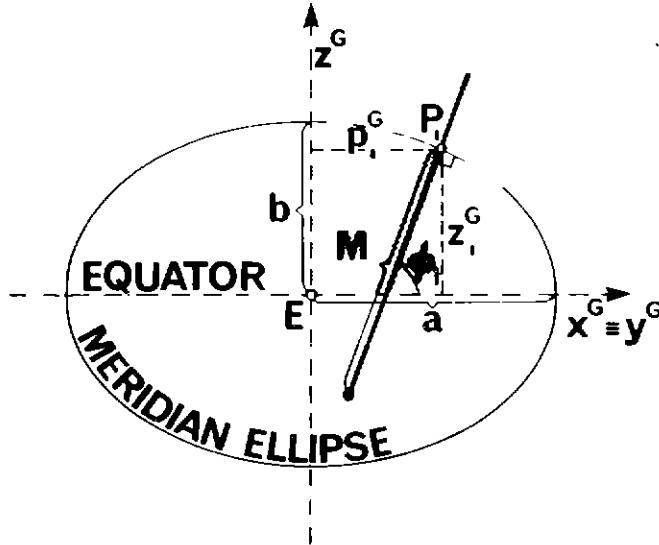


FIG. 7.14. Radius of curvature of a meridian ellipse.

By taking the total derivative of (8), and substituting for  $dp^G/dz^G$  into (9), one obtains

$$\tan \phi_i = \left( \frac{a}{b} \right)^2 \frac{z_i^G}{p_i^G}. \quad (7.10)$$

Denoting the eccentricity of the ellipse by  $e$ , i.e.,

$$e^2 = \frac{a^2 - b^2}{a^2}, \quad (7.11)$$

and evaluating  $z^G$  from (8), substitution in (10) yields

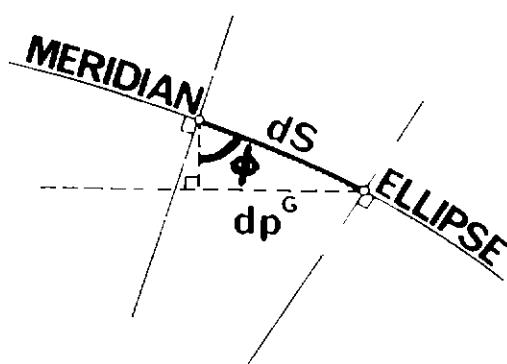
$$p_i^G = \sqrt{\frac{a^2 \cos^2 \phi_i}{1 - e^2 \sin^2 \phi_i}}. \quad (7.12)$$

The equation for  $M(\phi)$  can now be formulated. Since  $dS$  can be written as

$$dS = - \frac{dp^G}{\sin \phi} \quad (7.13)$$

(cf. FIG. 15), then after differentiation of (12) one gets

$$M(\phi) = \frac{a(1 - e^2)}{(1 - e^2 \sin^2 \phi)^{3/2}}. \quad (7.14)$$

FIG. 7.15. Relation between  $dP^G$  and  $dS$ .

Equation (6) finally becomes

$$S_{ij} = a(1 - e^2) \int_{\phi_i}^{\phi_j} (1 - e^2 \sin^2 \phi)^{-3/2} d\phi. \quad (7.15)$$

An identical equation can be written for the second meridian arc.

The two equations, relating the two meridian arcs to the two unknown parameters  $a, e$  (via the latitudes of the end points), contain elliptic integrals. These integrals can be evaluated only by using series expansion (see §3.2). The series are then inverted to obtain a solution for  $a$  and  $e$  in an iterative manner. Subsequently,  $f$  is computed from  $e$  using the following formula:

$$f = 1 - \sqrt{1 - e^2}, \quad (7.16)$$

which is derived from (4) and (11). The final results of the described French experiment, together with other results, are given in TABLE 2.

TABLE 7.2  
Biaxial (geocentric) ellipsoids

Solution	$a$ (km)	$f^{-1}$	Remarks
Eratosthenes [SCHWARZ, 1975]	5950	$\infty$	Sphere assumed
French experiment [BOHM, 1972]	6376 568	310 3	Basis for the definition of a metre
1924 International [HAYFORD, 1909]	6378 388	297 0	
1967 International [IAG, 1971]	6378 160	298 247	Also used by other international scientific bodies
Smithsonian [GAPOSHKIN, 1973]	6378 140	298 256	
U.S. Department of Defense [SEPPELIN, 1974]	6378 135	298 26	
1980 International [IAG 1980]	6378 137	298 257	Also used by other international scientific bodies

Determinations made in the late nineteenth and early twentieth centuries increasingly employed triangulation networks—i.e., geodetic horizontal networks, available at that time, with only horizontal angles and an occasional base line measured—instead of just meridian arcs. Conceptually, the use of networks is equivalent to the approach using meridian arcs with the difference being that one works with two dimensions on the ellipsoidal surface rather than one; astronomically determined longitudes must also be available for this technique.

When gravity data from various points on the surface of the earth became available, they could also be used in determining the shape of the earth. The following *Clairaut's theorem* describes the relation between gravity and the size and shape of the earth [HEISKANEN AND MORITZ, 1967]:

$$f = \frac{5}{2} \frac{\omega^2 a}{\gamma_E} - \frac{\gamma_P - \gamma_E}{\gamma_E}. \quad (7.17)$$

Here,  $\gamma_E$  is normal gravity on the equator,  $\gamma_P$  is normal gravity on the poles, and the quantity  $(\gamma_P - \gamma_E)/\gamma_E = \bar{f}$  is known as *gravity flattening*.

In the early twentieth century many biaxial (geocentric) ellipsoids were proposed. This led the IAG, in 1924, to recommend to its member countries that the ellipsoid determined by the American geodesist HAYFORD [1909] be adopted as an *international ellipsoid*. At the time, that ellipsoid (see TABLE 2) was considered to be the best fitting to the geoid. That ellipsoid was also meant to generate the normal gravity as given by (6.17), and thus a precedent was set for the dual rôle of all future geocentric ellipsoids.

Subsequent studies showed that the 1924 international ellipsoid could be regarded as only an approximation, from the point of view of its fit to the geoid. Consequently, a new international reference ellipsoid (see TABLE 2) was adopted by the IAG in 1967 [IAG, 1971]. Comparing the 1967 reference ellipsoid parameters with the best parameters currently available e.g., GAPOSHKIN'S [1973] or SEPPELIN'S [1974] (cf. TABLE 2), the decision was taken by IAG to adopt a new ellipsoid in 1979 [IAG, 1980].

On the other hand, the flattening of the international ellipsoid seems to have been chosen quite well. From a comparison with the latest results, it appears that the value of  $f^{-1} = 298.257$  is accurate to about  $10^{-2}$ , suggesting an accuracy of about  $10^{-7}$  in  $f$ . To be able to compare the accuracy in  $f$  with that of  $a$ , let us look at the change  $da$  in  $a$  implied by a change  $df$  in  $f$ . Differentiating (4), while holding  $b$  fixed, one finds

$$df = \frac{b}{a^2} da \doteq \frac{da}{a}. \quad (7.18)$$

When  $10^{-7}$  is substituted for  $df$ , the actual accuracy in  $f$  corresponds to an accuracy in  $a$  of under one metre. The reason why the flattening can be determined so much more accurately is that it can easily be derived from the perturbations of satellite orbits, as will be seen in Chapter 23. Ever since the first determination of  $f$  from

satellite observations (1 : 297.9, BUCHAR [1958]), its accuracy has been considerably better than that obtained from various terrestrial observations.

The length of the major semi-axis  $a$  can be derived from distances observed on the surface of the earth. Such a derivation provides a size of the earth that is scaled by the accepted value for the velocity of light. This is because all the distances used for geodetic purposes are now being measured with instruments based on registering the time lapse between emission and reception of a reflected electromagnetic wave. On the other hand, Chapters 22 and 23 will show that when gravity data or satellite orbits are used in deriving geoidal heights, then it is the adopted value of the earth's mass that scales the derived quantities. Therefore, in some problems, the accuracy depends to a large extent on the compatibility of the two adopted standards.

Of course, in theory, there is only one biaxial ellipsoid best fitting the geoid, in the least-squares sense. It is the one, defined by  $a$  and  $f$ , that satisfies the following condition (cf. §3.2):

$$\min_{a,f} \oint_S N^2 dS, \quad (7.19)$$

where  $N$  is the geoidal height referred to this ellipsoid, and  $S$  is the surface of the ellipsoid. It has been shown by HEISKANEN AND MORITZ [1967] that the enforcement of a similar condition, namely,

$$\min_f \oint_S (\xi^2 + \eta^2) dS, \quad (7.20)$$

where  $\xi$  and  $\eta$  are the components of the deflection of the vertical related to the same ellipsoid, leads to identical results for  $f$ , if  $a$  is held fixed. If, conversely,  $f$  is held fixed, the condition (20) yields an  $a$  compatible with that from (19). In practice, this best-fitting ellipsoid can only be approximately realized because of the lack of data.

So far, biaxial ellipsoids whose purpose is to approximate the geoid as closely as possible all over the earth's surface have been treated. These ellipsoids are usually assumed or forced to have their axes coincident with the principal axes of inertia of the earth. For this reason, they are called *geocentric ellipsoids*. There exists, however, another family of ellipsoids whose purpose is not to represent the earth's size and shape but to serve as a reference ellipsoid for geodetic positioning, i.e., as horizontal geodetic datums. These ellipsoids were introduced in §7.1, and they are mentioned again here for the sake of completeness.

These *geodetic reference ellipsoids* are normally selected so as to approximate the geoid only in a certain region (continent, group of countries, etc.) and do not meet the requirement of being geocentric. The size and shape of such an ellipsoid is generally chosen beforehand, and the closeness to the geoid is achieved by an appropriate positioning of the ellipsoid within the earth. FIG. 16 illustrates the relation between the two kinds of ellipsoids.

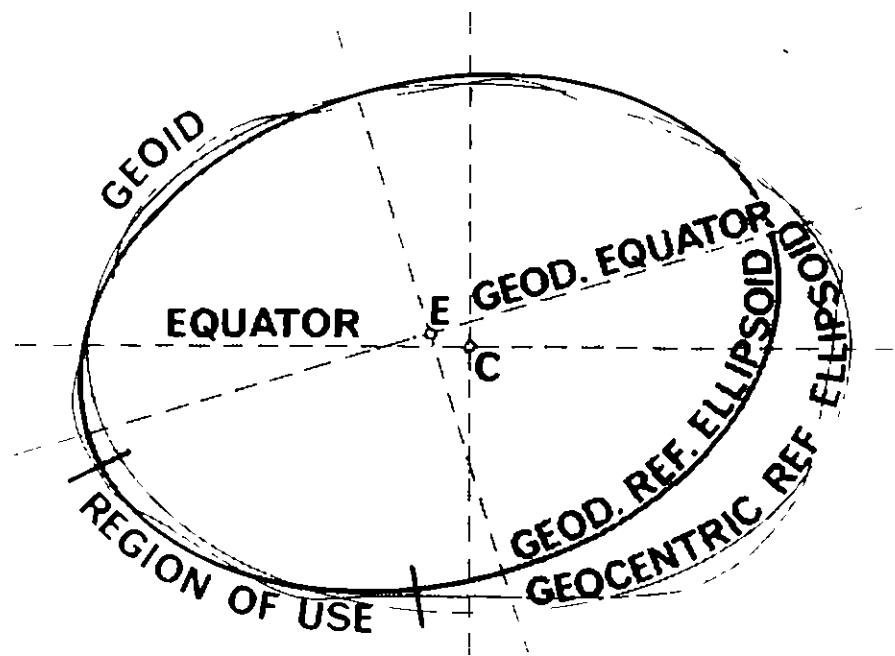


FIG. 7.16. Geocentric and geodetic ellipsoids.

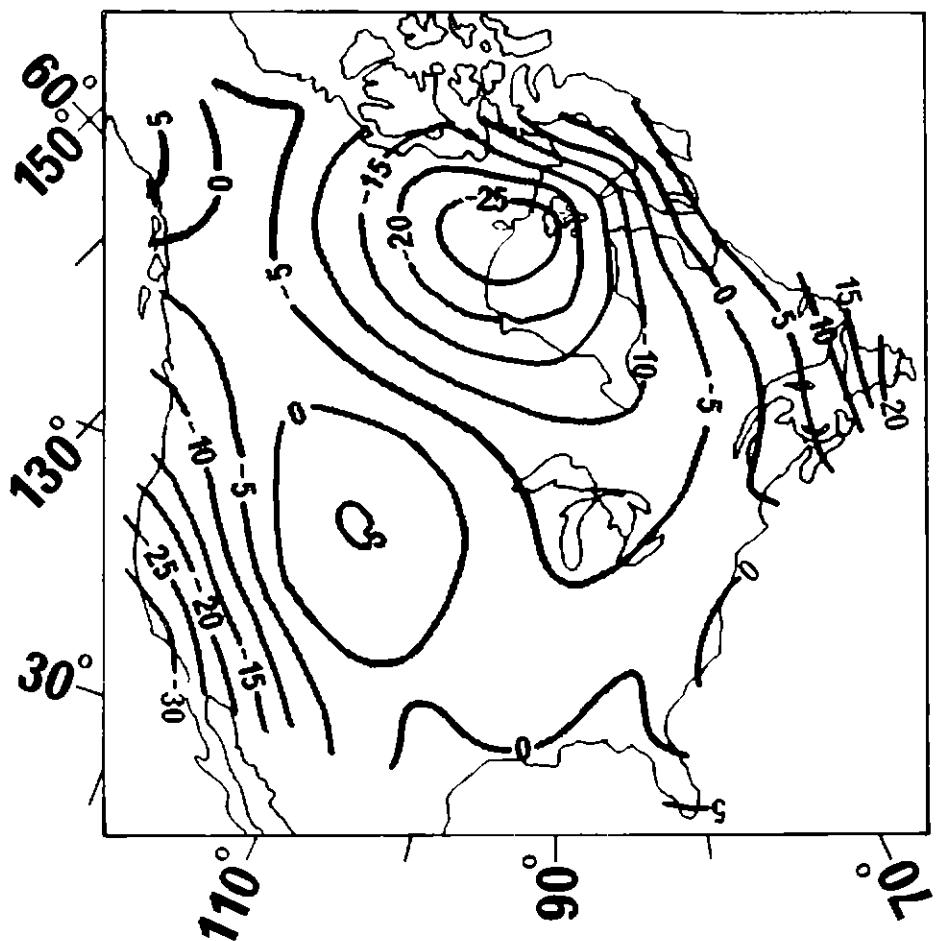


FIG. 7.17. The geoid as referred to the NAD 27. Contours in metres.

The reason why the geodetic datum has to closely approximate the geoid in the region where it is used will be explained in §18.3. Because only regional fit is required from a geodetic datum, its fit can be much better than that of the geocentric ellipsoid. This point is shown on FIG. 17, where a patch of geoidal surface in North America referred to the North American geodetic datum (NAD 27) is depicted [VANÍČEK AND MERRY, 1973]. The reader should compare this with the corresponding part of the geoid related to a geocentric ellipsoid (FIG. 6.16). The regions of use of the main datums were shown in FIG. 3. Scores of other datums are used in other parts of the world.

To conclude this section, let us point out that for some purposes, in disciplines like astronomy or cartography, the figure of the earth can be adequately represented as a *reference sphere*. As shown in the previous section, the departure of the best-fitting sphere from the best-fitting ellipsoid is about 10.7 km on the poles and on the equator.

#### 7.4. Other mathematical figures of the earth

Modern geodetic theories employ two additional surfaces (figures of the earth) not discussed thus far. They will be briefly described here; for a comprehensive treatment, the reader is referred to Chapter 22.

The surface designed to approximate the physical surface of the earth is the *telluroid*. The telluroid is defined as the surface whose height above a geocentric reference ellipsoid is the same as the height of the terrain above the geoid [HIRVONEN, 1960]. FIG. 18 shows the relationship between the telluroid, terrain, geoid, and geocentric ellipsoid.

A close relative of the geoid is a surface called the *quasigeoid*. It was introduced by MOLODENSKIJ ET AL. [1960] as the solution to the practical problems encountered in geoidal computations. When computing the geoidal height from terrestrial data, and also when computing heights above the geoid, one has to assume some distribution of masses within the upper layers of the earth. This assumption weakens the confidence in the results. The quasigeoid, on the other hand, can be derived without any such assumption (see §22.2). In other words, the quasigeoidal height, usually

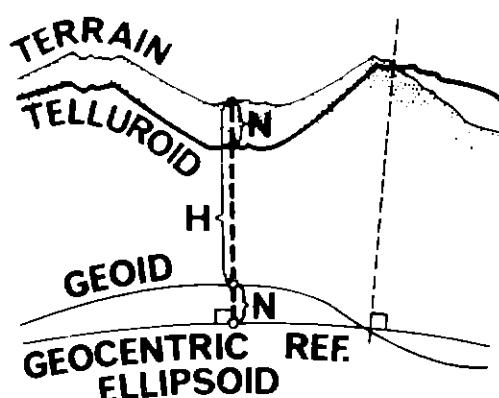


FIG. 7.18. Telluroid.

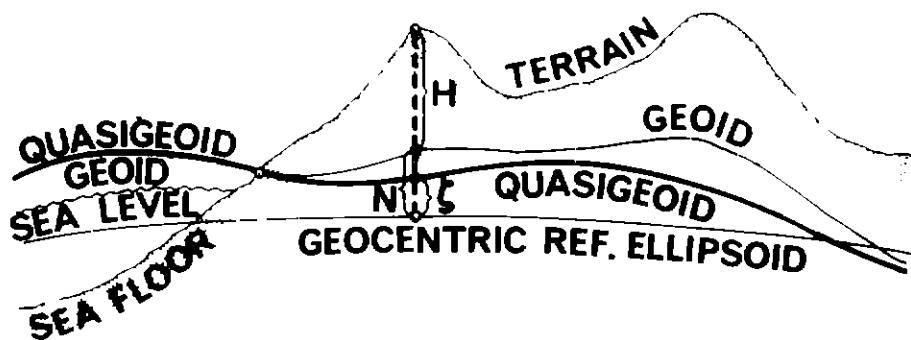


FIG. 7.19. Quasigeoid.

called the *height anomaly* and denoted by  $\xi$  (see FIG. 19), can theoretically be computed exactly. In Molodenskij's system, the telluroid is defined as the surface whose height above the geocentric ellipsoid is identical to that of the physical surface above the quasigeoid. It can be shown that the normal potential  $U$  on that telluroid is equal to the actual potential  $W$  at the corresponding point on the earth's surface.

The evident disadvantage of the quasigeoid, compared with the geoid, is that it does not have any physical meaning. It is purely a mathematical creation and is not an equipotential surface of the earth's gravity field. However, there is no disadvantage in using the quasigeoid as a reference surface for heights, as will be seen in §19.1. Heights referred to the quasigeoid are now used in the U.S.S.R. and most Eastern European countries.

The quasigeoid coincides with the geoid on open seas, where the geoidal height  $N$  equals the height anomaly  $\xi$  (cf. FIG. 19). Inland, these two surfaces depart by as much as a few metres [ARNOLD, 1960]. The difference  $N - \xi$  is highly correlated with the height of terrain, and thus the largest differences occur under mountains.

For geophysical purposes, another figure sometimes used to represent the idealized earth is a biaxial ellipsoid, with flattening slightly smaller than that of the best-fitting ellipsoid. This is the *hydrostatic equilibrium ellipsoid*. In order to understand its proper role, let us first return to the best-fitting geocentric ellipsoid and its (normal) gravity field. As shown in §6.4, this gravity field is designed to have one of its equipotential surfaces coincident with the ellipsoid itself. If the earth were fluid and laterally homogeneous, the shape of the earth's surface in hydrostatic equilibrium would be the equipotential ellipsoid. From various indications it is known that the earth's core is fluid, the mantle is weak, and a thin layer of solid crust is on the top. Hence its overall behaviour, under permanent stress induced by gravity, should be close to that of a highly viscous fluid (cf. §6.3).

The density distribution  $\sigma$  and the flattening  $f$  of any stratum of such a fluid body in hydrostatic equilibrium are related through *Clairaut's equation* [MELCHIOR, 1972]:

$$\frac{d^2f}{dr^2} + \frac{2\sigma r^2}{\int_0^r \sigma s^2 ds} \frac{df}{dr} + \left( \frac{2\sigma r}{\int_0^r \sigma s^2 ds} - \frac{6}{r^2} \right) f = 0, \quad (7.21)$$

where  $r$  is the mean radius of the stratum. LEDERSTEGER [1967] and others maintain

that there exists no reasonable, stratified, density distribution, compatible with other geophysical data, that would satisfy (21) for the observed flattening 1 : 298.25. This opinion is disputed by MORITZ [1973] and other scholars.

JEFFREYS [1963] approached this problem from the other end. He computed the *hydrostatic flattening* of an earth model, with the same physical characteristics as the real earth, and came to the conclusion that its value should be 1:299.67. Similar values, e.g., 1:299.8 [HENRIKSEN, 1960], are obtained from the dynamic flattening  $H$  (cf. §5.3; do not confuse with height) derived from observations of the earth's precession. The dynamic flattening is related to the geometrical flattening  $f$  by the following equation [MELCHIOR, 1972]:

$$H = \frac{f - \frac{1}{2}m}{1 - \frac{2}{5}\sqrt{\frac{5m}{2f} - 1}}, \quad (7.22)$$

where

$$m = \frac{\omega^2 a^3}{GM}, \quad (7.23)$$

which permits a direct evaluation of  $f$  from  $H$ , if  $m$ , the *geodetic parameter*, is known.

These results show that the actual mass distribution and the actual flattening are not compatible with the assumption of hydrostatic equilibrium and the lateral homogeneity of masses within the strata. An explanation for the disagreement between the two values of the flattening was sought in the failure of the earth to adjust itself to the present, slower spin velocity. The observed, exaggerated flattening would thus be the remnant of the faster spin the earth has experienced in the past. However, the viscosity, necessary to sustain such a retardation in adjustment, seems to be too high compared with the values obtained from other geophysical considerations. More probably, the larger, actual flattening corresponds to the actual distribution of masses thus indicating a concentration of denser masses in the equatorial belt of the mantle, which might be explained by the centrifugal effect. The discrepancy between the observed and the hydrostatic equilibrium flattenings is of the order of 100 metres. This is comparable to the deviation of the geoid from symmetry in the equatorial plane or the pear-shape effect (cf. §7.2), as pointed out by GOLDREICH AND TOOMRE [1969].

KAULA AND O'KEEFE [1963] suggest that, for the purpose of geophysical interpretations of the earth's gravity field, gravity anomalies and the geoid be referred to this ellipsoid rather than the best-fitting one. The rationale is that, in this case, the long wavelength meridian features that reflect the difference in the theoretical and observed flattenings, better depict the mass anomalies within the earth. To illustrate this point, the two representations of the geoid, according to GAPOSHKIN [1973], are given in FIGS. 20 and 21.

TABLE 3 recapitulates the order of magnitude of the maximum departure of various pertinent pairs of surfaces encountered in this chapter.

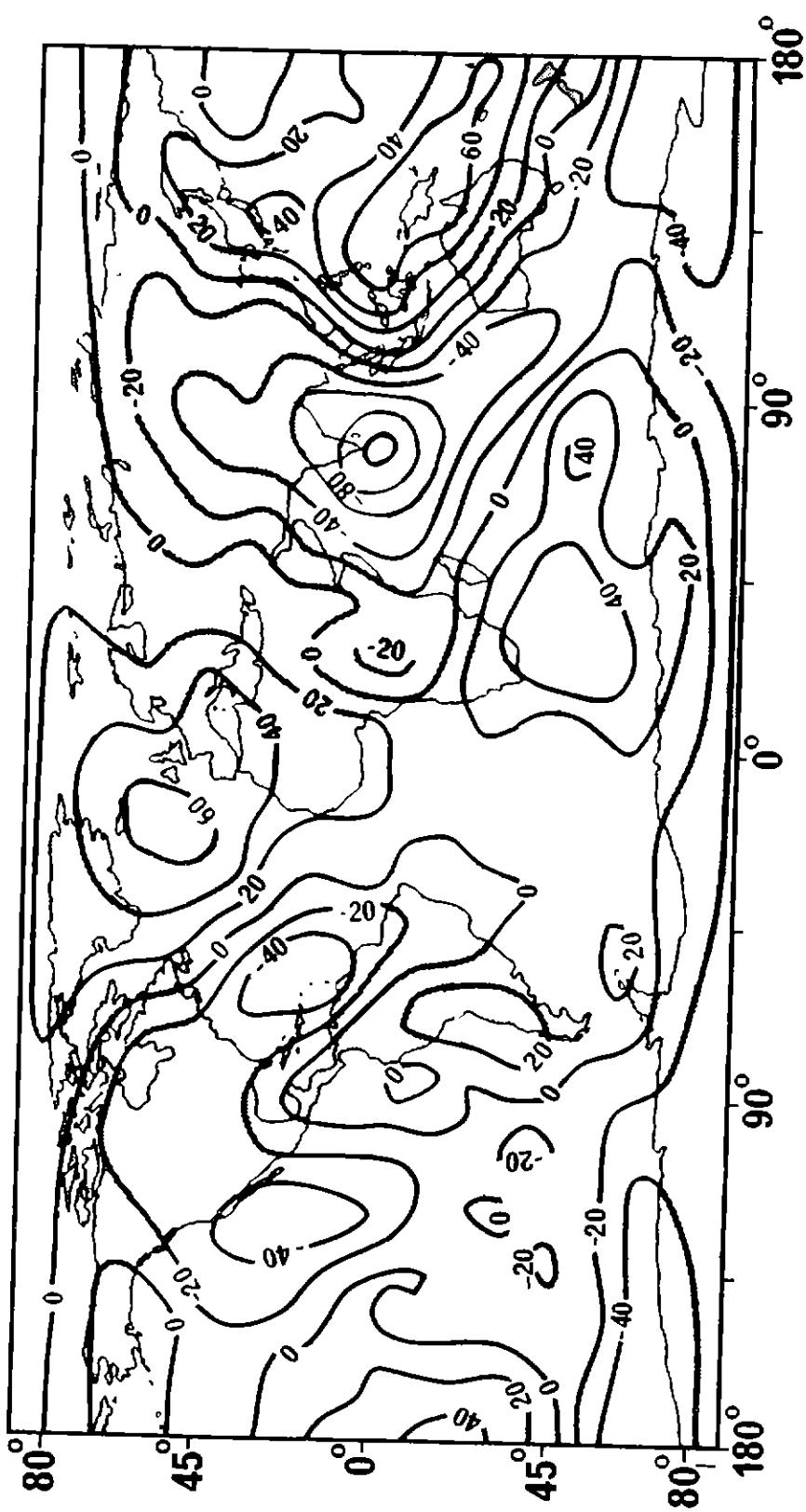


FIG. 7.20. Global geoid referred to the best-fitting ellipsoid with  $f=1:298.256$ . Contours in metres.

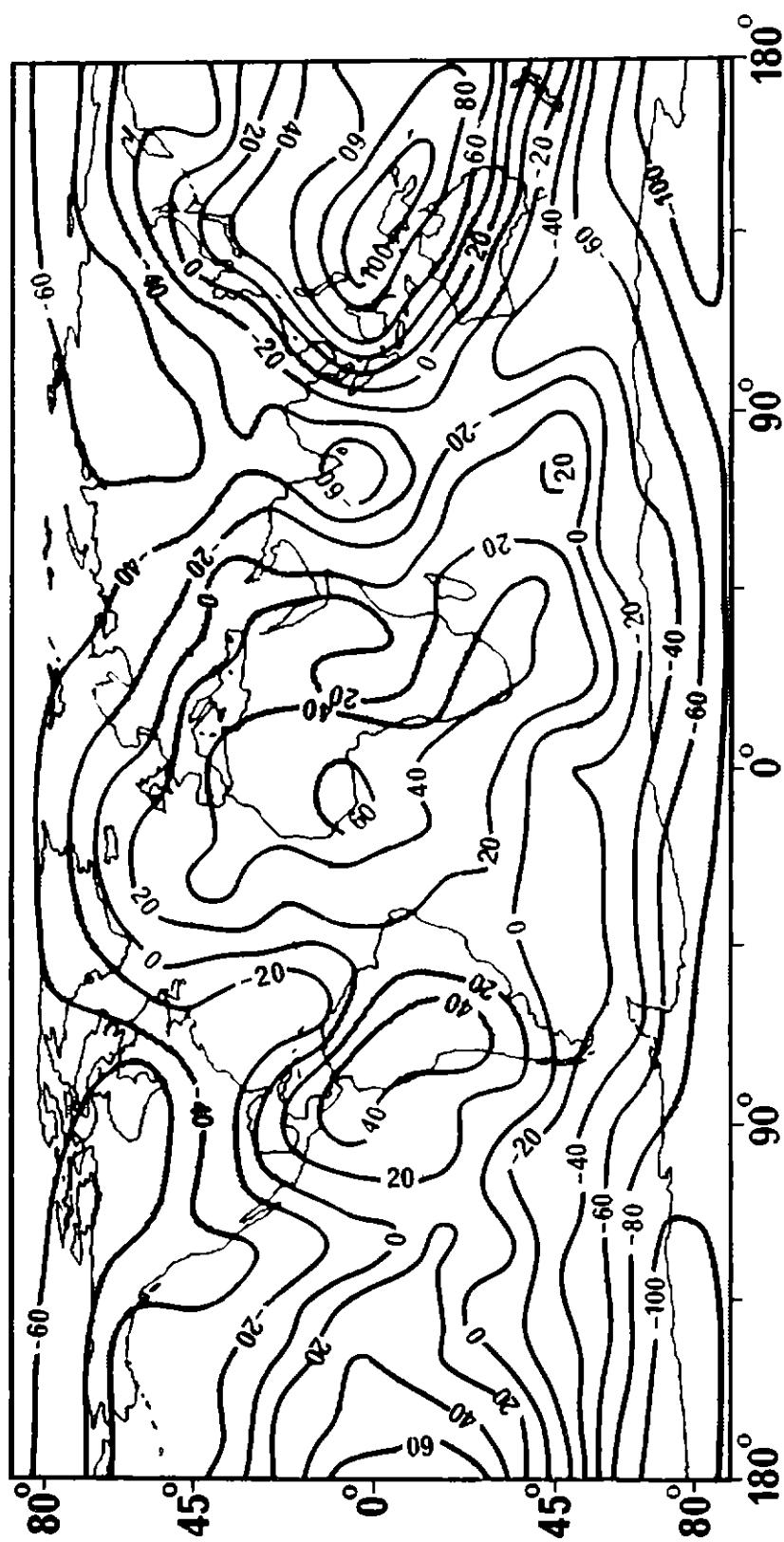


FIG. 7.21. Global geoid referred to the ellipsoid with hydrostatic flattening  $f = 1 : 299.67$ . Contours in metres.

TABLE 7.3  
Departures of pertinent pairs of surfaces

Surfaces	Order of magnitude of maximum departure (metres)
terrain-geoid/mean sea level	$10^4$
mean sea level-geoid/quasigeoid	1
geoid-quasigeoid	1
geoid-ellipsoid } telluroid-terrain }	$10^2$
biaxial-triaxial ellipsoids	$10^2$
ellipsoid-sphere	$10^4$

To conclude this chapter, let us take an overall look at the surfaces purporting to portray the figure of the earth. On the one hand, the determination of the shape of the terrain—and its discretisation, the networks—is one of the ultimate aims of geodesy. This surface and its point representation are the most complex, and thus the most awkward, to work with. On the other hand, disregarding the sphere, the biaxial ellipsoids are the simplest surfaces. These are used as reference surfaces in both positioning (the horizontal geodetic datums) and in the study of the earth's gravity field (the geocentric ellipsoid).

Somewhere in between, playing almost the role of the terrain on the seas and the role of vertical geodetic datums inland, are the geoid and quasigeoid. They are also somewhere in between as far as their complexity is concerned. The remaining surfaces, i.e., the sphere, triaxial ellipsoids, hydrostatic equilibrium ellipsoids, analytical representations of the terrain, spheroids, and telluroids, are used only for special purposes. The last two will be treated more fully in Part V.

## CHAPTER 8

### EARTH AND ITS DEFORMATIONS IN TIME

One of the main aims of geodesy, as seen in Chapter 6, is the determination of the earth's shape. Because this shape varies in time, locally as well as globally, the geodesist finds it necessary to account for the temporal deformations. In Chapter 5, it was seen how the spin of the earth varies with time in both the direction and velocity. In Chapter 6, reference was made to the time variations of the gravity field. This chapter makes a systematic survey of the dynamic phenomena that change the shape of the earth and, consequently, change the positions of points on it.

From the point of view of time dependence, these variations are classified as secular (linear, slow, creeping), periodic (with periods ranging from fractions of a second up to tens of years), and episodic (suddenly accelerating and decelerating). Because a systematic study of the secular and long-periodic phenomena was only begun a few decades ago, a qualitative and quantitative understanding of most of them has not yet been achieved. Therefore it is often not known if some of them are secular or long-periodic in nature.

At the other end of the spectrum are the high-frequency phenomena (seismic and other tremors), which normally have little effect on geodetic work. It is difficult, however, to off-handedly dismiss the importance of such phenomena in geodesy. A certain kind of movement considered negligible in one context may prove significant in another context. It can be said, however, that normally our main interest is in secular and low-frequency movements with periods of a few hours and longer. In addition, usually only recent (i.e., movements that have occurred in the past century or so) and contemporary movements are of interest to geodesists. This is in contrast to other earth scientists who may study the history of movements that occurred in the past thousands or millions of years.

To understand better how the earth's shape changes, first consider what its response to deforming forces would be if the earth were fluid. It is easily seen that in this case the earth's surface would respond in the same way as the oceans' surface does. If, on the other hand, the earth were rigid, there would be no deformation whatsoever. The response of the real earth is somewhere between the two extremes of being fluid and rigid. In addition, if the deforming force is applied only for a short duration, or if it changes very rapidly, then the deformation of the earth is elastic. This means that as soon as the force is removed, the recovery of the original shape is practically instantaneous. On the other hand, if the force is of a very long duration, the response is viscous: after cessation of the force, the recovery of the

original shape is only gradual. A medium that behaves in such a fashion (response dependent on the frequency of deforming force) is known as visco-elastic.

Some of the phenomena about to be described are better understood than others. Of the significant effects, the longest known and best understood are the tidal, which will be dealt with in the first section. Crustal loading phenomena and isostasy are introduced in the next section. Tectonic movements, i.e., the movement of the tectonic plates of the earth's crust, have become a focus of intensive research in the past two decades and are the topic of the third section. Assembled in the final section are the remaining movements originating from, e.g., various meteorological changes and man-made causes—just to mention a few.

### 8.1. Tidal phenomena

By *tidal deformation* we mean phenomena caused by the variations in the gravitational force exerted by celestial bodies; the corresponding force is known as the *tidal force*. At any point within, or on the surface of, the earth, the gravitational force exerted by a celestial body (say the moon) can be split into two components: first, that equalling the gravitational force acting at the centre of mass of the earth, and second, that equalling the remainder. This can be seen in FIG. 1. The first part is the force governing the motion of the earth as a whole, as discussed in Chapter 5. The second part, shown by bold arrows, is the tidal force. It is interesting to note that the tidal force at *D* acts in the opposite direction to the celestial body. This is caused by the earth accelerating towards the attracting body at the same rate as its centre of mass *C*, except that the near side (*A*) is accelerating more and the far side (*D*) is accelerating less than the centre of mass. As a whole, it can be seen from FIG.

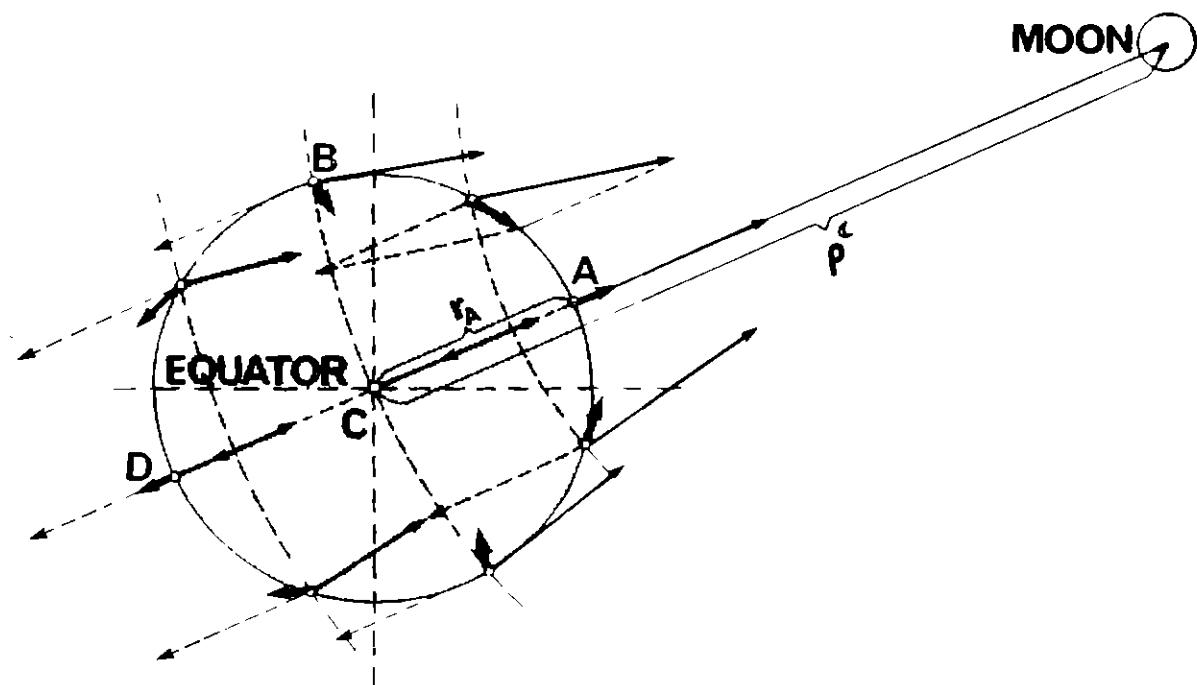


FIG. 8.1. Tidal force due to the moon.

1 that the tidal force tries to deform the equipotential surfaces so that their shape prolates in the direction of the celestial body. More precisely, the shapes will be elongated in the direction of the resultant force exerted by the configuration of all the celestial bodies.

Let us concentrate initially on the most significant body—the moon. That part of the *tidal acceleration* which is due to the moon—the lunar tidal acceleration—is denoted by  $\bar{a}_t^{\mathbb{C}}$ . Through the use of the law of universal attraction (6.1), the following equation can be written for the absolute value of  $\bar{a}_t^{\mathbb{C}}$  at point  $A$ , located on the surface of the earth and on the line connecting  $C$  with the moon (cf. FIG. 1):

$$a_t^{\mathbb{C}}(A) = \frac{GM^{\mathbb{C}}}{(\rho^{\mathbb{C}} - r_A)^2} - \frac{GM^{\mathbb{C}}}{(\rho^{\mathbb{C}})^2} = \frac{GM^{\mathbb{C}}}{(\rho^{\mathbb{C}})^2} \left[ \left(1 - \frac{r_A}{\rho^{\mathbb{C}}}\right)^{-2} - 1 \right]. \quad (8.1)$$

The mass of the moon is denoted by  $M^{\mathbb{C}}$ ,  $\rho^{\mathbb{C}}$  is the distance between the centres of mass of the earth and the moon, and  $r_A$  is the distance of  $A$  from the earth's centre of mass. If  $M^{\mathbb{C}} \doteq 7.38 \times 10^{25}$  g, an average  $\rho^{\mathbb{C}} \doteq 3.84 \times 10^{10}$  cm [GODIN, 1972], and  $r_A \doteq 6.371 \times 10^8$  cm, then  $a_t^{\mathbb{C}}(A) \doteq 0.111$  mGal is obtained. Because the total acceleration  $a^{\mathbb{C}}$  exerted by the moon on the earth is about 3.3 mGal, the tidal acceleration represents at most 3.4% of the total acceleration. A similar simplified estimate for point  $B$  (cf. FIG. 1) yields  $a_t^{\mathbb{C}}(B) \doteq 0.055$  mGal.

Again, as in Chapter 6, it is more convenient to work with potentials than with accelerations. Because the tidal acceleration is defined as the acceleration corresponding to the difference of two forces (one equalling that acting at the centre of mass of the earth and the other acting at the point of interest), due to the linearity of the gradient operation it can be expressed as a gradient of the difference of the potentials of the two corresponding accelerations, i.e., as the gradient of *tidal potential*. However, to be able to formulate the two potentials, the two accelerations have to be considered as vector fields. One is the radial attraction acceleration field  $\bar{a}^{\mathbb{C}}$  of the moon, and the other is a constant acceleration field  $\bar{c}^{\mathbb{C}}$  defined by the following identity:

$$\bar{c}^{\mathbb{C}} = \bar{a}^{\mathbb{C}}(C). \quad (8.2)$$

To begin with, let us formulate the first potential. According to FIG. 2, the whole lunar potential at an arbitrary point  $A$  is given by (cf. Chapter 6)

$$W^{\mathbb{C}}(A) = \frac{GM^{\mathbb{C}}}{\rho_A^{\mathbb{C}}} = \frac{GM^{\mathbb{C}}}{((\rho^{\mathbb{C}})^2 + r_A^2 - 2r_A\rho^{\mathbb{C}} \cos Z_A^{\mathbb{C}})^{1/2}}, \quad (8.3)$$

where  $Z_A^{\mathbb{C}}$  is the zenith distance (angle) of the moon at  $A$ . When the denominator is developed into a series of Legendre's functions  $P_n(\cos Z_A^{\mathbb{C}})$  (see §3.2), an expression is obtained that is easier to handle:

$$W^{\mathbb{C}}(A) = \frac{GM^{\mathbb{C}}}{\rho^{\mathbb{C}}} \sum_{n=0}^{\infty} \left( \frac{r_A}{\rho^{\mathbb{C}}} \right)^n P_n(\cos Z_A^{\mathbb{C}}). \quad (8.4)$$

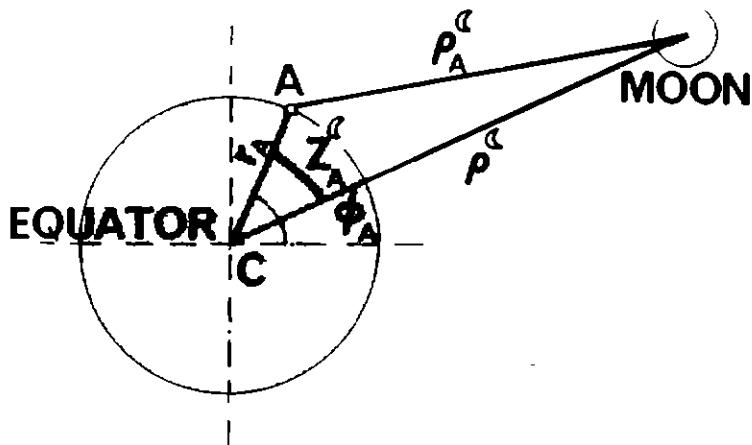


FIG. 8.2. Tidal potential.

The second potential is that of the constant field  $\bar{c}^{\mathbb{C}}$ . The reader can verify (by direct computation of  $\bar{a}_t = \nabla V^{\mathbb{C}}$ ) that this potential, at an arbitrary point  $A$ , is given by

$$V^{\mathbb{C}}(A) = \frac{GM^{\mathbb{C}}}{\rho^{\mathbb{C}}} + \frac{GM^{\mathbb{C}}}{(\rho^{\mathbb{C}})^2} r_A \cos Z_A^{\mathbb{C}} = \frac{GM^{\mathbb{C}}}{\rho^{\mathbb{C}}} \sum_{n=0}^1 \left( \frac{r_A}{\rho^{\mathbb{C}}} \right)^n P_n(\cos Z_A^{\mathbb{C}}). \quad (8.5)$$

The following formula for the lunar tidal potential is derived by taking the difference of these two potentials:

$$W_t^{\mathbb{C}}(A) = W^{\mathbb{C}}(A) - V^{\mathbb{C}}(A) = \frac{GM^{\mathbb{C}}}{\rho^{\mathbb{C}}} \sum_{n=2}^{\infty} \left( \frac{r_A}{\rho^{\mathbb{C}}} \right)^n P_n(\cos Z_A^{\mathbb{C}}). \quad (8.6)$$

An identical expression can be written for the solar tidal potential  $W_t^{\odot}$  simply by replacing the lunar mass  $M^{\mathbb{C}}$  by the solar mass  $M^{\odot}$ , the distance of the moon  $\rho^{\mathbb{C}}$  by the distance of the sun  $\rho^{\odot}$ , and the zenith distance  $Z^{\mathbb{C}}$  by  $Z^{\odot}$ . From astronomical observations it was learned that, on average, the solar potential is about 46% of the lunar potential. Contributions from other celestial bodies are much smaller; listed in TABLE 1 are the principal components in descending order. Normally, only the *luni-solar tidal potential*,  $W_t = W_t^{\mathbb{C}} + W_t^{\odot}$ , is considered. The smaller the ratio  $r/\rho$ , the faster the series (6) converges. For practical use, to an accuracy of a few percent, it suffices to take only the first terms for the moon and the sun. The addition of the second lunar term ( $W_3^{\mathbb{C}}$ ) improves the accuracy to about 0.03%.

TABLE 8.1

Relative contributions to tidal potential from different celestial bodies

Body	Tidal potential
Moon	1.0
Sun	0.4618
Venus	0.000054
Jupiter	0.0000059
Mars	0.0000010

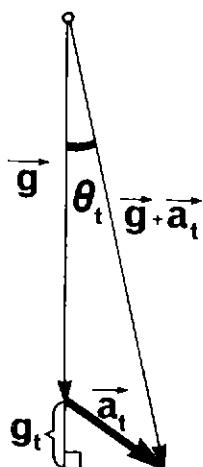


FIG. 8.3. Tidal tilt.

The tidal potential can easily be converted into the three kinds of tidal deformations of particular interest to us here. The *tidal gravity variation*,  $g_t$  (i.e., the vertical component of  $\vec{a}_t$ ), is simply evaluated as the negative radial derivative of the potential. *Tidal tilt*,  $\theta_t$ , of equipotential surfaces is related to tidal gravity, as shown on FIG. 3. Finally, the *tidal uplift* of equipotential surfaces,  $u_t$ , is obtained from the tidal potential through a formula parallel to (6.30). A more detailed treatment of these three effects will be given in §25.1.

The luni-solar potential, at any point in and on the earth, obviously varies with time because the distances  $\rho^{\text{C}}$ ,  $\rho^{\text{O}}$  and zenith distances  $Z^{\text{C}}$ ,  $Z^{\text{O}}$  vary with time. The main periodicities of these variations, i.e., those with the largest amplitudes, are semidiurnal and diurnal. The origin of the diurnal period band is easily understood from the motion of the moon and the sun. The presence of the semidiurnal band is understandable when it is realized that the tidal potential is the same whether the celestial body stands overhead or under the observer. This can be seen from the symmetry of the forces on FIG. 1; the forces would look practically the same if the moon stood on the left, instead of on the right, of the earth. The predominant contribution to the tidal potential is lunar semidiurnal, denoted in the literature by  $M_2$ , with the period of half a lunar day, i.e., half a revolution of the earth with respect to the moon.

Apart from the two main frequency bands, the tidal potential displays other periodicities which reflect the periodicities in the various parameters of the lunar and earth orbits and their interactions. FIG. 4 shows the maximum relative contributions of the main *tidal frequencies* (tidal constituents) to the complete luni-solar potential [GODIN, 1972]. Some of the most important constituents are denoted by capital letters with subscripts: 1 denotes diurnal, and 2 denotes semidiurnal. Generally,  $M$  denotes lunar and  $S$  solar constituents; note the already mentioned provenience of the  $M_2$  contribution. Other letters have been chosen for reasons considered outside our interest.

The tidal potential, in addition to being a function of time, is also a function of position, i.e., the position of the observer affects the magnitude of the observed tidal

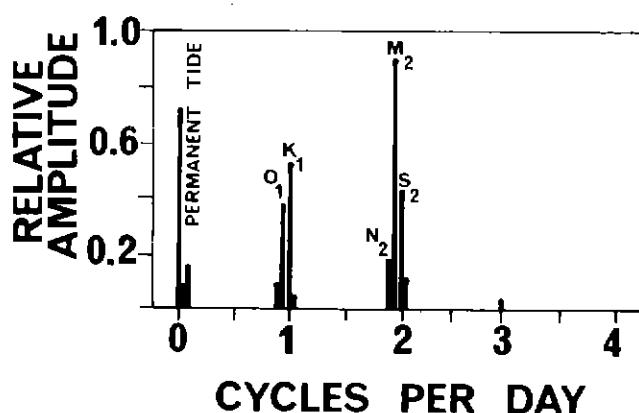


FIG. 8.4. Maximum relative amplitudes of contributions of tidal frequencies.

potential and thus all the tidal phenomena. Because the rate of the earth's spin defines the rate with which the parameters  $\rho$  and  $Z$  vary in (6), the longitude of the observer influences only the timing, i.e., the phase of the individual constituents. The effect of latitude is more complicated: it affects the mean value of the angle  $Z$  (cf. FIG. 2) and thus influences the magnitude of the potential disparately for different frequencies. The relative amplitudes of the individual tidal waves, shown in FIG. 4, become modified by latitude according to their frequencies. For instance, the amplitude of the diurnal waves is largest at latitude  $\phi = 45^\circ$  and zero for the equator and the poles. Semidiurnal waves are strongest at latitude  $\phi = 0^\circ$  and disappear at the poles. A more detailed treatment is given in §25.1.

It should be noted that the *permanent tidal uplift*, i.e., tide with zero frequency, is responsible for an increment in the permanent flattening of the earth's equipotential surfaces. It contributes a depression of 28 cm at the poles and an uplift of 14 cm along the equator [MELCHIOR, 1966], which amounts to a decrement of 0.006 in  $f^{-1}$ . The relative amplitude contribution of the permanent tide is about 80% of that of  $M_2$  (cf. FIG. 4).

Theoretically,  $u_1$  is also the amount of *water tide*. As explained in Chapters 6 and 7, disregarding for the moment its inhomogeneity, water is constantly adjusting its level to coincide with an equipotential surface and, therefore, follows the movements of the equipotential surface in response to tidal force. In nature this is seldom the case. Because the bottom slopes toward the shore, the incoming tidal water is trapped and piles up along the shoreline. Only small islands, rising steeply from the sea floor, are free from this effect. This piling effect evidently is more pronounced in confined and shallow areas, like bays and straits, where the water tide can be magnified several times. The Bay of Fundy (Canada) and the Bristol Channel and Liverpool Bay (England) are notable for their sea tide well in excess of 10 metres. The largest tidal range in the world (16.3 m) has been recorded in Minas Basin (Bay of Fundy, Canada) [MCWHIRTER AND MCWHIRTER, 1975].

In these confined areas, other dynamic phenomena can also be observed. *Resonance* and *non-linear interference* effects, peculiar to the locality, can cause sizeable amplitude and phase distortion of the individual tidal waves. These effects are strongly frequency dependent; therefore, the amplitude relations among the ob-

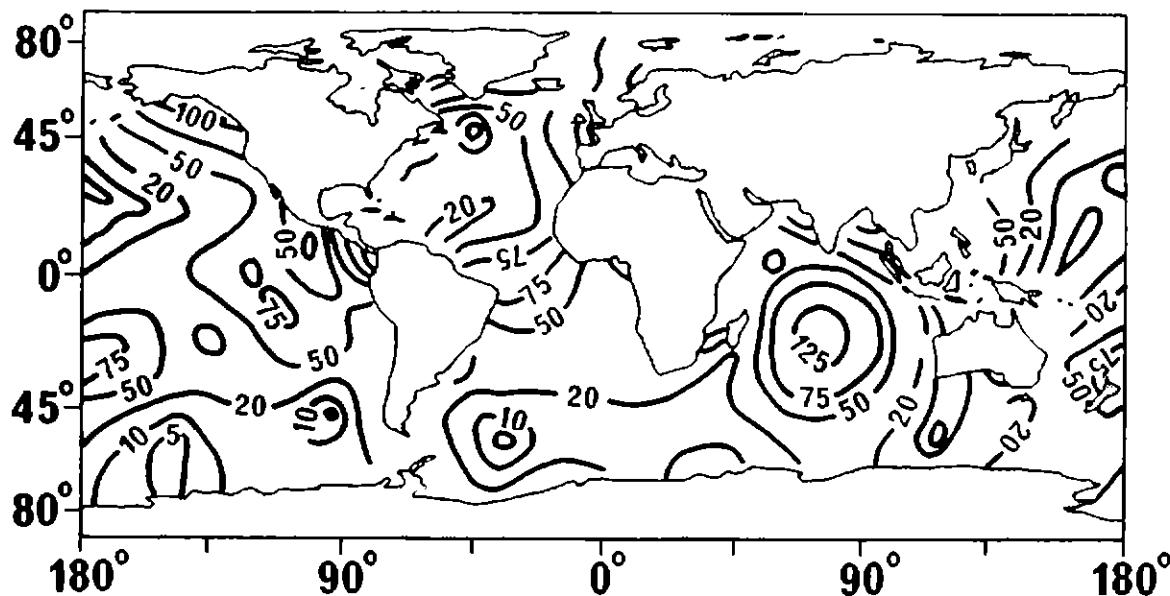


FIG. 8.5. Corange chart for  $M_2$  tide amplitude. Contours in centimetres.

served tidal waves in these localities are generally quite different from those of the theoretical, so-called equilibrium, waves. Exploration of these questions is beyond the scope of this book; the interested reader is referred to specialized oceanographical literature, e.g., DOODSON [1957], HILL [1966], and MACMILLAN [1966].

The sea tides have been, and are being, recorded at thousands of tide stations along various coasts. This makes it possible to study the global behaviour of the actual sea tide from the acquired data. Many such studies have been conducted, and here the results are shown of one of the more recent determinations [HENDERSHOTT, 1972]: FIG. 5 presents the *corange chart* for the  $M_2$  sea tide. The dynamic distortions of this tidal wave are clearly visible, even though the coastal regions are not charted. Similar charts showing the time of occurrence of the maximum of the tidal wave are known as *cotidal charts*. The last tidal phenomenon to be mentioned is the atmospheric tide. Since Chapter 9 is devoted to the earth's atmosphere, this phenomenon will be treated there.

As we know by now, the earth is not rigid. Therefore, not only the earth's gravity field but also the shape of the earth's body changes under the action of the tidal force. Although the details involving the response of the deformable earth are going to be treated in Chapter 25, here, in TABLE 2, at least the ranges of the maximum actually observable deformations are given.

TABLE 8.2  
Maximum actual ranges of observed tidal deformations

Type of deformation	Maximum range
observed gravity variation ( $g_t$ )	0.28 mGal
relative tilt of the earth's surface ( $\theta_t$ )	0.017"
relative uplift of equipotential surface ( $u_t$ )	53 cm

In closing, it should be pointed out that the tidal changes of the earth's body also cause the earth's surface to deform. Thus the distances on the surface of the earth contract and expand; and the angles of intersecting lines undergo changes. The relative changes are fairly small—of the order of  $10^{-8}$  [MELCHIOR, 1978]—and undetectable in routine geodetic operations. The distance variations, however, are being observed by other means as will be shown in §25.2.

## 8.2. Crustal loading deformations

According to our current knowledge, the earth's crust is composed of plates of lighter, solidified material, of an average density of  $\sigma \doteq 2.67 \text{ g/cm}^3$  [HEISKANEN AND VENING MEINESZ, 1958], floating on denser matter ( $\sigma \doteq 3.27 \text{ g/cm}^3$ ) that is weakened by partial melting resulting from heat and pressure. It is rather difficult to distinguish exactly where the solid crust ends and the weakened mantle begins. The two sources of information, seismology and rheology, do not agree completely; thus, depending on the criterion chosen, the boundaries are perceived in different depths [OFFICER, 1974]. There is a tendency to use the term *crust* only for the top 10 to 30 km thick layer [RUNCORN, 1967] and to refer to the solid plates as the *lithosphere*. The lithospheric plates are known from rheological investigations to vary in thickness between, say, 10 and 80 kilometres. The uppermost part of the mantle, to a depth of some 300 to 400 km, is called the *asthenosphere*—see FIG. 6. For a more detailed treatment, the reader is referred to, e.g., GASS ET AL. [1972] and OFFICER [1974].

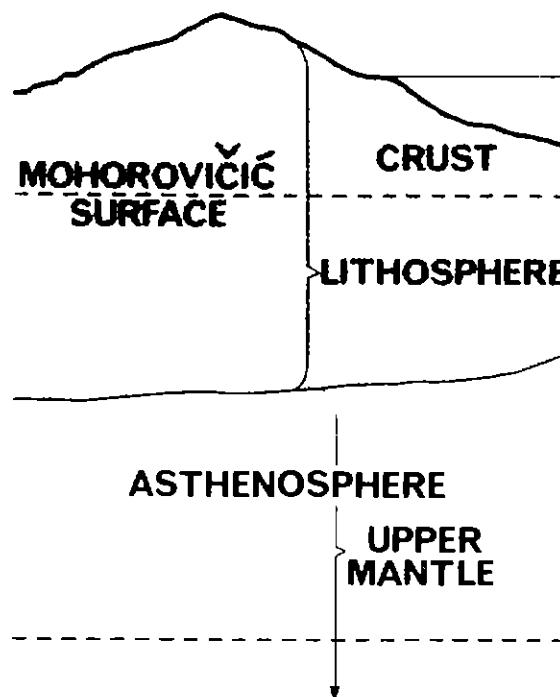


FIG. 8.6. The upper layers of the earth.

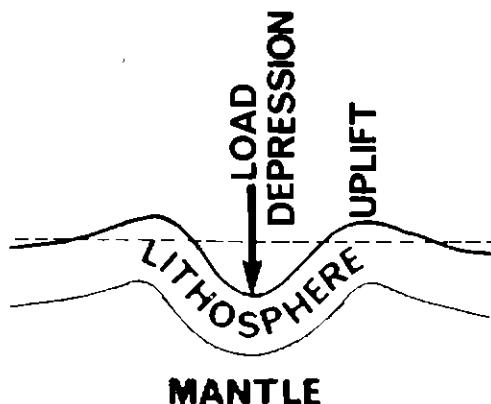


FIG. 8.7. Response to load.

These plates are subjected to loads arising from different phenomena that take place on the earth's surface. Any such load produces regional, vertical deformation of the crust. It should be clear that a load at one point on the surface of the earth will cause the crust to yield not only immediately underneath the load but also in the surrounding area because of the lateral strength of the lithosphere. The subsidence will be the largest immediately under the load and will gradually diminish with distance from the load. To maintain the same volume (except for the volume expended on elastic compression) of the earth, the depression is accompanied by an uplift in peripheral regions (cf. FIG. 7). The relationship between the amount of subsidence and the distance from the load depends on the rheology of the lithosphere and the mantle as well as on the size of the load. The elastic response will be treated in detail in §25.3.

Turning now to the existing sources of load, in descending order of their importance, the most conspicuous is *ice*. To comprehend the immenseness of the ice load, at present there is an estimated  $3 \times 10^7 \text{ km}^3$  of ice covering the Antarctic. This ice represents a total load of  $2.7 \times 10^{19} \text{ kg}$  compared with the weight of ice covering Greenland, estimated to amount to  $3 \times 10^{18} \text{ kg}$  [WALCOTT, 1975]. In connection with the earth's surface movements, the most pertinent ice load is that which covered large parts of Canada, Fennoscandia, Siberia, the Himalayas, the Alps, and the southern tip of South America during the last glaciation. This glaciation is thought to have ended some 6000 to 10000 years ago. The ice sheets reached a maximum thickness of a few kilometres at the peak of the ice age. It has been estimated that the vertical depression of the crust amounted to about 500 m in the centre of the Northern Hemisphere's largest glaciated areas. Approximately the same amount is estimated for Greenland today.

An equally important source of loading is the water produced by the melting of the ice described above and called, in brief, the *ice melt*. The overall weight of this water is the same as that of the melted ice and thought to be of the order of  $3 \times 10^{19} \text{ kg}$ ; but the water spreads over a much larger area than that covered by the glaciers. If the area covered by the sea is taken to be about  $3.7 \times 10^8 \text{ km}^2$ , and if the water from the melting ice had been distributed uniformly (which it was not, as will be

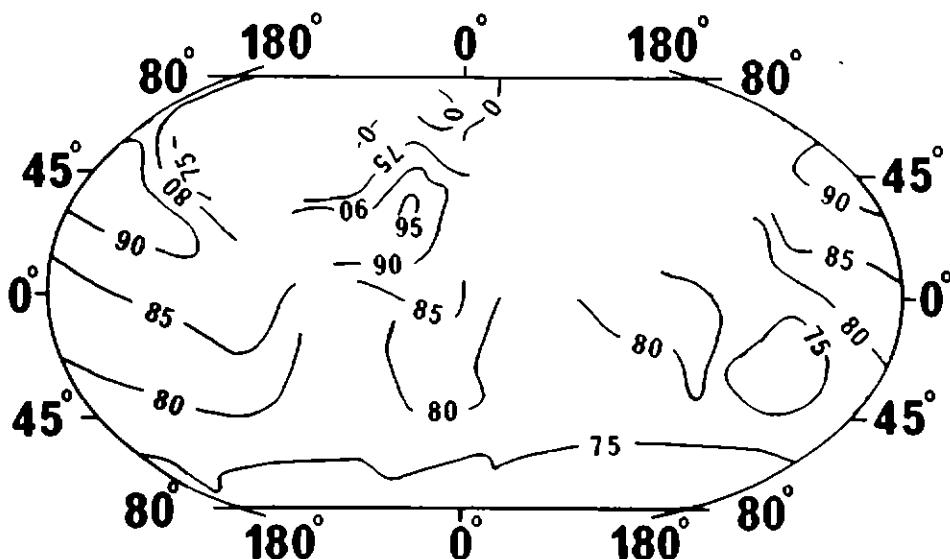


FIG. 8.8. Global rise of sea level relative to depressed crust. Contours in metres.

seen later), the sea would have risen by some 80 metres. PELTIER ET AL. [1978] investigated the response of the earth to loading by water formed by the ice melting after the last glaciation. FIG. 8 shows their estimates of the global rise of sea level, relative to the depressed crust, that occurred from the onset of melting.

*Deposits* by large rivers of solid particles in sedimentary basins are another source of load. The Mississippi River (U.S.A.), a well-documented example, deposits about  $2 \times 10^{11}$  kg of silt annually. This increases to an estimated  $8 \times 10^{11}$  kg during the years of large floods [MUELLER, 1975]. Recent subsidence of the order of 10 cm has been reported around the estuary. Of course, sedimentation has been responsible for the growth of very many sedimentary basins, the world over, during the geological times,

As already seen in the previous section, a significant loading source is the *tidal water*. A typical, say semidiurnal, tidal wave 5 m high over an area of  $10^4$  km $^2$  represents a considerable load of  $5 \times 10^{13}$  kilograms. The induced pressure of  $5 \times 10^2$  g/cm $^2$  is, however, of less importance. Because of the (tidal) frequency of the load, the response can be considered elastic and, as such, is usually modelled. Let it suffice here to show, for illustration, the elastic deformation of the crust in response to the  $M_2$  component of sea tide loading, obtained for the area of the Bay of Fundy, by integrating the load from cotidal charts (FIG. 9). This particular deformation is going to be treated in detail in §25.3; it happens to be one of the few kinds of deformations that is predictable with a degree of certainty.

Of similar significance are loads induced by large *water reservoirs*. One of the world's largest artificial reservoirs [MCWHIRTER AND MCWHIRTER, 1975]—Lake Kariba on the Zambezi River (Africa)—stores about  $1.5 \times 10^{14}$  kg of water over an area of 6650 km $^2$  [GOUGH AND GOUGH, 1970]. A brief computation shows that the pressure is about  $2 \times 10^3$  grams per centimetres squared. Depression of up to 13 cm was observed after the lake had been filled. How much of this deformation, though, is due to soil compaction (consolidation) is difficult to know. A specific load, similar

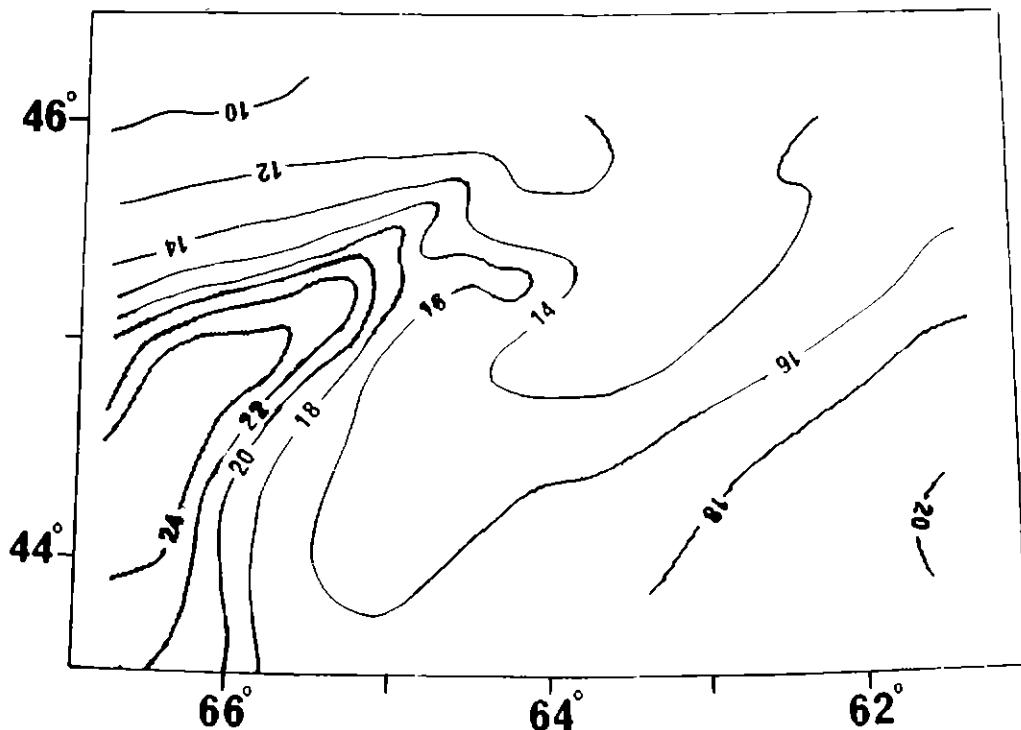


FIG. 8.9. Vertical displacement due to  $M_2$  sea tide load. Contours in millimetres. (Courtesy of the Earth Physics Branch, DEPARTMENT OF ENERGY, MINES AND RESOURCES [1977b], Ottawa, Canada.)

to the annual rate of a deposit load, can be obtained from *large cities*. Other man-made objects are too light to warrant consideration. From natural causes, *magma* from erupting volcanoes should be named; little is known, however, about its loading effect.

Finally, for comparison, let us mention high *barometric pressure* systems. Their (elastic) loading effect, measurable only with very sensitive instruments, is of the same magnitude as that of abundant *precipitation* and is about two orders of magnitude smaller than the sedimentation effect. Accumulation of *snow* may provide load of at least one order of magnitude larger than rain water.

To understand the crustal rebound after a visco-elastic deformation has taken place and the load has been removed, it is necessary to first outline the theory of the static equilibrium of the earth's crust—the principle of *isostasy*. If the solid lithospheric plates float on weak asthenospheric material in equilibrium, the variations in the depth of submersion must be balanced out by the variations in lithospheric density and thickness (including the topographical relief). This state of equilibrium is what the lithosphere strives for after having been depressed by the load that has subsequently been removed. There exist three main working hypotheses which model the required relationship between the crustal thickness and density.

PRATT's [1855] model assumes the boundary between the lithosphere and asthenosphere to be flat, i.e., the depth of this boundary below the sea level to be uniform. For the crust to be in equilibrium, those parts which rise must have a lower density ( $\sigma$ ) and vice versa. To be able to compute the appropriate density, one imagines the lithosphere to consist of independent blocks as shown on FIG. 10. The blocks must

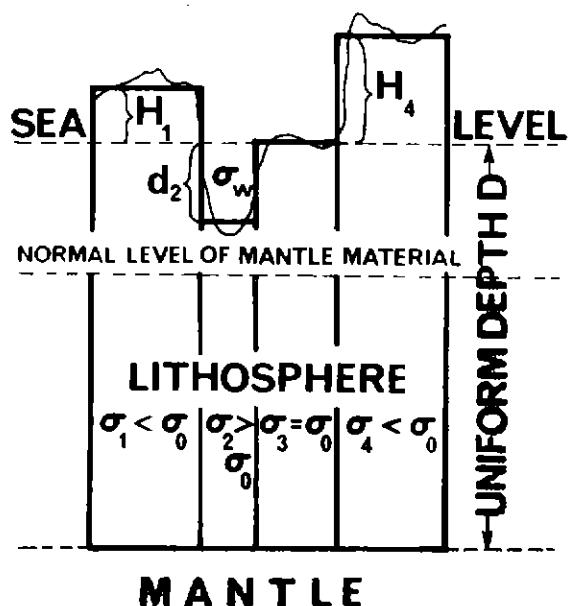


FIG. 8.10. Pratt's model.

exert the same pressure on the mantle at uniform depth  $D$  to attain an equilibrium. From this condition, the density of the *continental lithosphere* as a function of the mean height  $H_i$  of the block above the sea level is obtained;

$$\sigma_i = \sigma_0 \frac{D}{D + H_i}. \quad (8.7)$$

A similar equation

$$\sigma_i = \frac{\sigma_0 D - \sigma_w d_i}{D - d_i}, \quad (8.8)$$

where  $\sigma_w \doteq 1.027 \text{ g/cm}^3$  is the density of ocean water, relates the density of the *oceanic lithosphere* to its average depth  $d_i$ . Through the use of a reasonable normal depth  $D = 30 \text{ km}$  and normal density  $\sigma_0 = 2.67 \text{ g/cm}^3$ , densities ranging between  $2.06$  and  $3.76 \text{ g/cm}^3$  are obtained. This range appears to be too large to be reconcilable with geological information.

On the contrary, AIRY's [1855] model does not allow for density variations but treats the lithosphere as having a variable depth. To maintain the equilibrium, the lithosphere must be thicker underneath a higher topographical relief and thinner underneath the oceans. For computational reasons, the lithosphere is again envisaged as composed of independent blocks.

With  $S$  denoting the normal depth of submersion into the mantle material, and using Archimedes's law, the following can be written for the departures  $R_i$  of actual depth from the normal depth  $D$  of the lithosphere (cf. FIG. 11),

$$\begin{aligned} \sigma_m S &= \sigma_0 D, \\ \sigma_m (S + R_i) &= \sigma_0 (D + H_i + R_i), \\ \sigma_m (S - R'_i) &= \sigma_0 (D - d_i - R'_i) + \sigma_w d_i, \end{aligned} \quad (8.9)$$

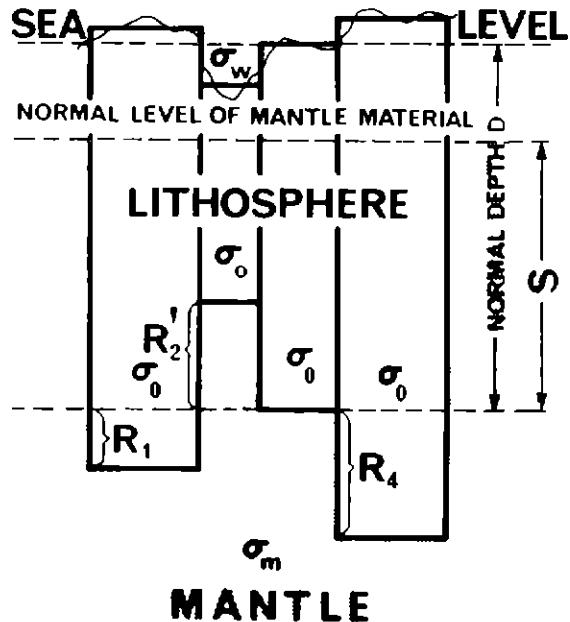


FIG. 8.11 Airy's model

where  $\sigma_m$  is the density of the upper mantle. Then the equations for the *roots of the continental blocks* are easily obtained:

$$R_i = \frac{\sigma_0}{\sigma_m - \sigma_0} H_i. \quad (8.10)$$

Similarly, the *anti-roots of oceanic blocks* are given by

$$R'_i = \frac{\sigma_0 - \sigma_w}{\sigma_m - \sigma_0} d_i. \quad (8.11)$$

Substitution in the above equations of values for  $\sigma_0$ ,  $\sigma_w$ ,  $\sigma_m$  yields

$$R_i \doteq 4.45 H_i, \quad R'_i \doteq 2.73 d_i. \quad (8.12)$$

If the normal depth is considered to be around 30 km, the calculated lithospheric depth agrees fairly well with the depth determined by seismology. However, the necessity of imagining the lithosphere to be broken into independently floating blocks causes some anxiety. In reality, the lithosphere is mostly continuous with the exception of the boundaries of a few large blocks, as will be seen in the next section.

This consideration led VENING MEINESZ [1931] to his modification of Airy's model. In his model, Vening Meinesz assumes that the blocks are stuck together and, therefore, respond as a continuous elastic layer to the load exerted by topographical relief. This means that the sinkage of the lithosphere in the mantle is distributed over a larger, compensating region, as seen in FIG. 12.

From the physical point of view, neither of the aforementioned hypotheses is completely satisfactory. From various sources, it is known that both the lithospheric density and thickness vary. Also, the lithosphere behaves as an elastic layer in some regions but is broken in others. For unbroken lithospheric plates, modern visco-elastic

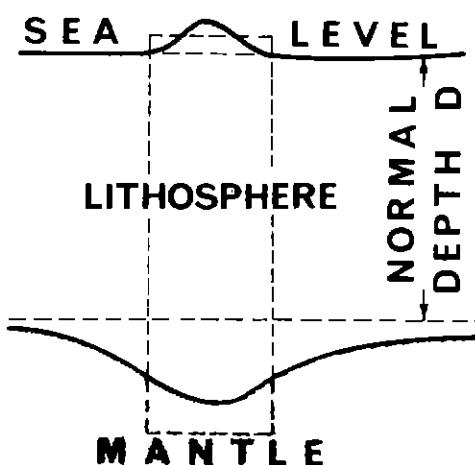


FIG. 8.12 Venning Meinesz's model

models (see, e.g., PELTIER ET AL. [1978]) seem to be quite successful in explaining the observed behaviour.

Let us now discuss the main known rebound phenomena. When the ice melted, the elastic relaxation was instantaneous. The non-elastic part of the deformation persisted; the lithosphere found itself and remained in a state of non-equilibrium. Since that time the crust has been rebounding, due to its buoyancy, to attain the isostatic equilibrium. This process is known as the *postglacial isostatic rebound*.

The speed of this rebound is controlled mainly by the viscosity of the upper mantle material: it has been shown theoretically that the speed decreases with time,

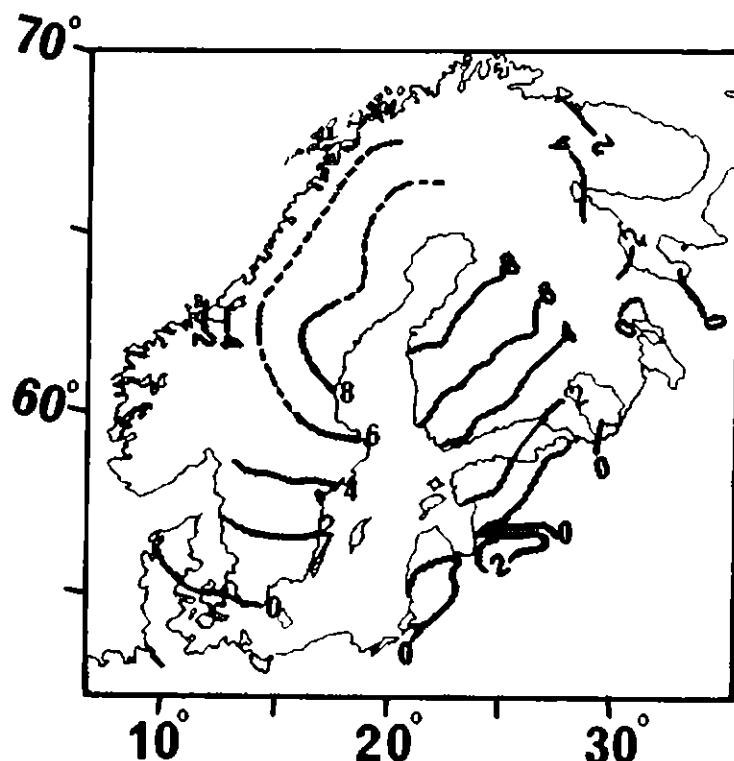


FIG. 8.13. Postglacial rebound in Fennoscandia. Contours in millimetres per year.

and it is thought that, by now, about three-quarters of the original deformation has recovered [HEISKANEN AND VENING MEINESZ, 1958]. FIG. 13 portrays the presently observed rate of postglacial rebound in Fennoscandia, as determined by KUKKAMÄKI [1975]. There are indications of comparable rates in the region of the Hudson Bay, Canada [WALCOTT, 1972]. Let us just mention here that the postglacial rebound is accompanied by a subsidence in peripheral regions; this phenomenon mirrors that of the glacial deformation, as depicted in FIG. 7.

Removal of any significant load produces an uplift. The well-known example of this is the rebound observed around the Great Salt Lake, Utah, U.S.A. FIG. 14 shows the actually observed vertical displacement after the *evaporation* of water weighing some  $8.2 \times 10^{15}$  kg, of an average depth of 330 m, according to GILBERT [1890]. *Erosion* of material in exposed areas, over an extended period of time, produces a comparable effect, although no actual examples of this effect were available to the authors.

Changes in gravity accompany the already discussed changes in the mass distribution. To obtain an estimate of the magnitude of these changes, let us examine the

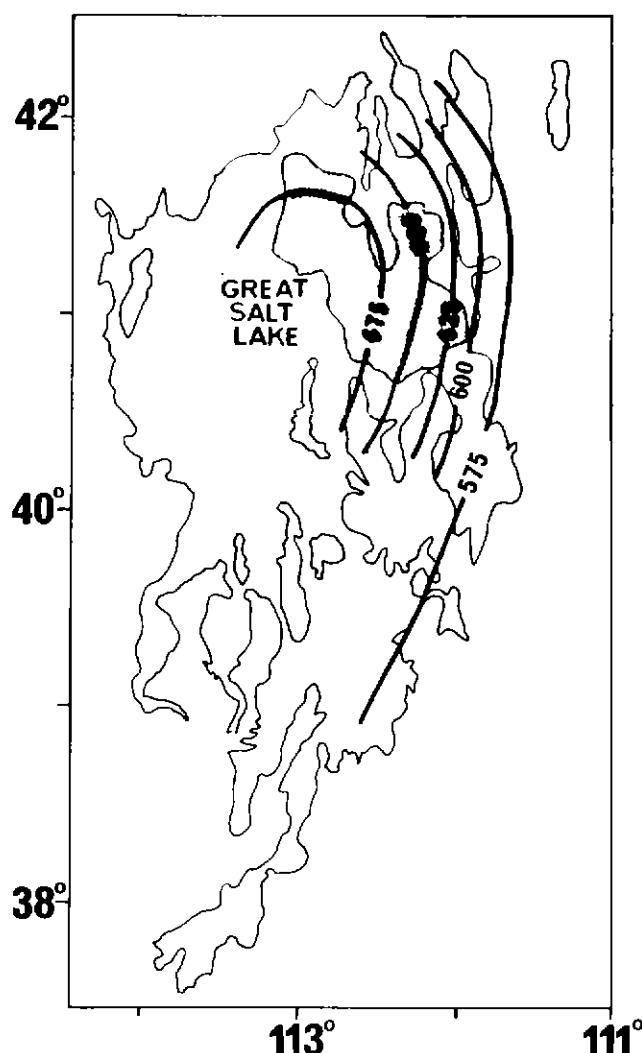


FIG. 8.14. Lake Bonneville rebound. Contours in centimetres.

postglacial rebound—the most significant of the presently observable loading phenomena. The Hudson Bay area is characterized by an average anomaly of about 30 mGal (cf. FIGS. 6.8 and 6.9), which is assumed to be mainly due to the remaining glacial depression. Accepting  $5 \times 10^3$  years as a likely time span for the disappearance of most of the remaining deformation, an estimate is obtained for the average annual gravity change of the order of  $6 \mu\text{Gal}$ .

For completeness, it must be said that crustal loading deformations also induce changes in horizontal distances and angles. These are, however, very minute and do not affect geodetic work.

### 8.3. Tectonic deformations

As seen in the previous section, the lithosphere, broken into plates, floats on the upper mantle material. Although the idea of lithospheric plates (with continents and ocean basins on them) drifting around on the upper mantle was proposed as early as at the beginning of this century [WEGENER, 1929], it was not until recently that it

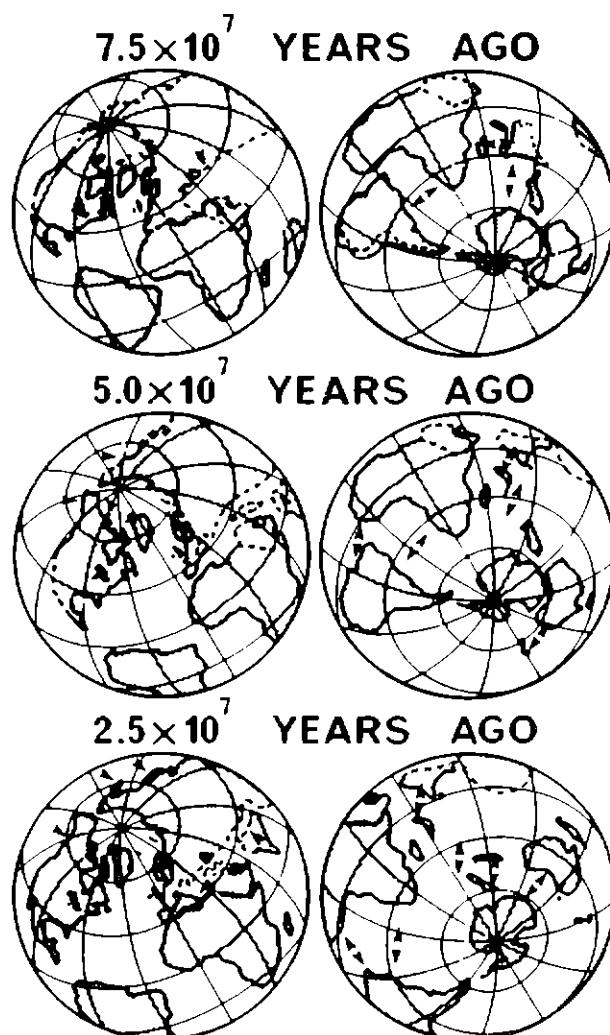


FIG. 8.15. Changes in past continental configurations.

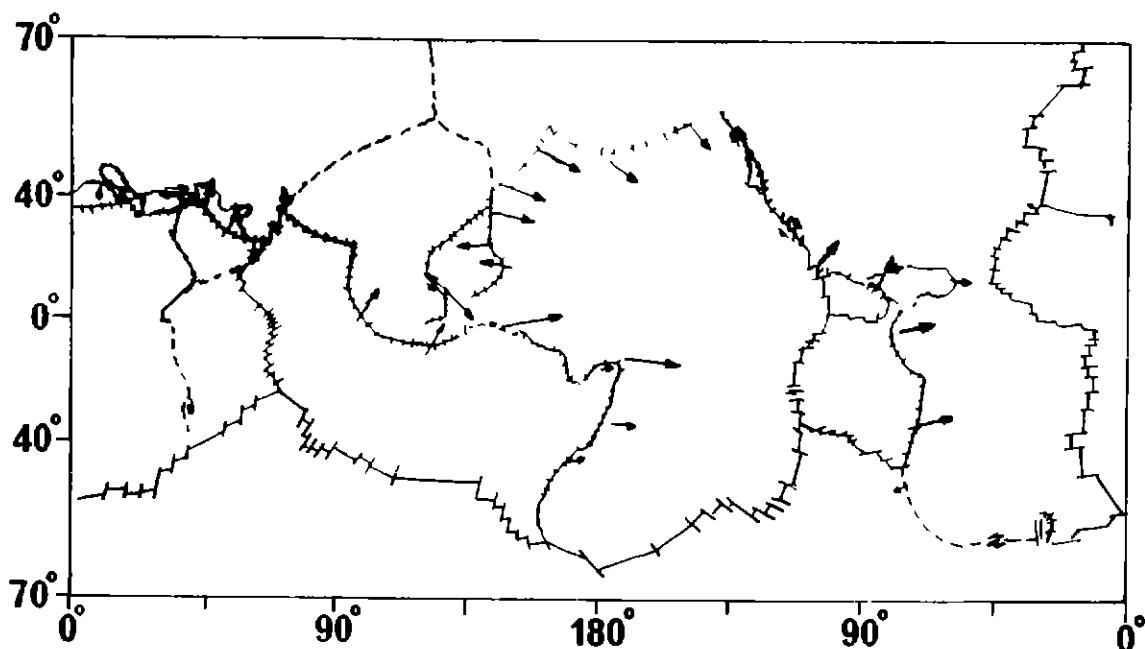


FIG. 8.16. Plate boundaries and motion.

received serious scientific consideration. The notion of lithospheric plate movement now seems to be firmly established, and vigorous research is under way to determine the plates' relative velocities, to explain the driving mechanism for the plate movement, and to delineate the exact plate boundaries.

Continental configuration 75 million, 50 million, and 25 million years ago, as reconstructed by IRVING [1977], is shown in FIG. 15. The boundaries of the major plates are now reasonably well known. FIG. 16 shows their present outline, according to LEPICHON ET AL. [1973]. Shown on the same figure are some estimates of the relative motion velocities ranging from zero to 1.1 cm per year in the south-west Atlantic and to 14.5 cm per year round New Guinea.

Present understanding of the driving forces of the plates is still inadequate. It appears probable that thermal convection within the asthenosphere is, in one form

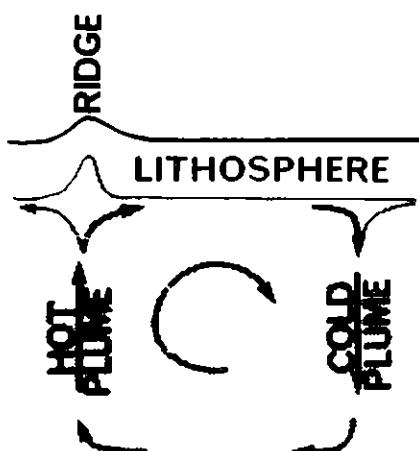


FIG. 8.17. Mantle convection cell.

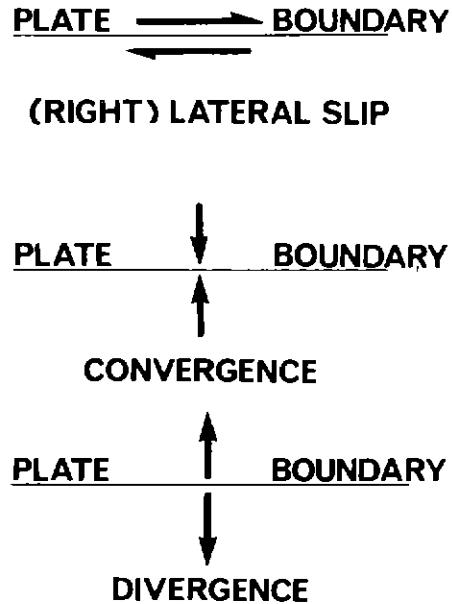


FIG. 8.18. Types of plate contact.

or another, certainly a contributor to the motion. FIG. 17 depicts one such possible concept [GAAS ET AL., 1972].

As the plates move around, three distinctly different kinds of plate boundary contacts are evident: lateral slip, convergence, and divergence, as shown on FIG. 18. In reality, combinations of slip with thrust and spreading also occur.

A diverging or *spreading boundary* is characterized by an opening in the crust where the mantle material rises, hardens, and creates new lithosphere. This phenomenon is accompanied by strong volcanic activity that produces ridges typical for these boundaries. An example of a spreading boundary is the mid-Atlantic Ridge—probably the most studied spreading boundary. The average rate of spreading of this ridge is about 2 cm per year [COULOMB, 1972].

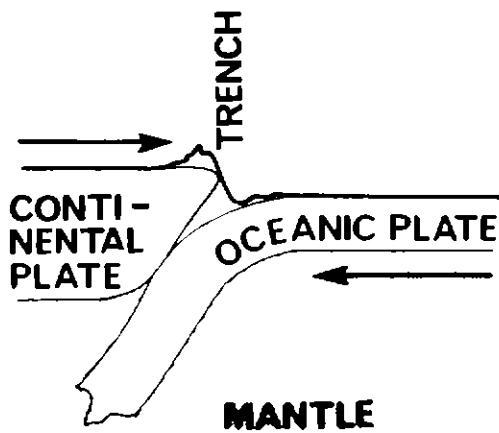


FIG. 8.19. Consuming plate boundary.

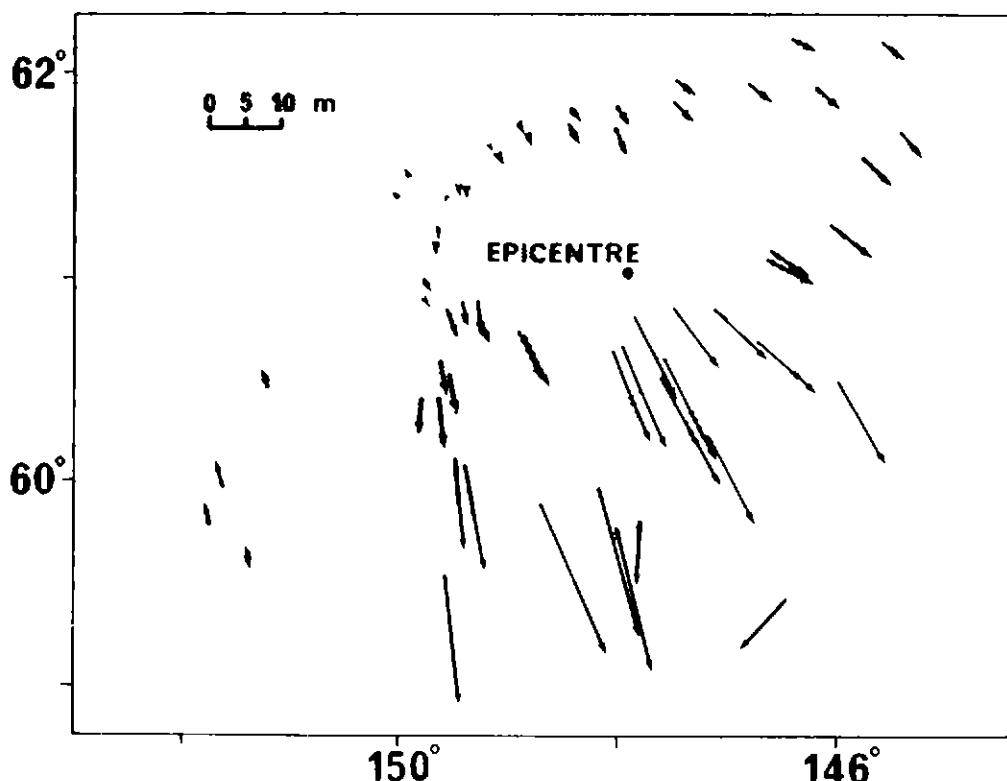


FIG. 8.20. Horizontal displacement resulting from the 1964 Alaska (U.S.A.) earthquake.

At the *converging boundary*, where two plates of different type thrust against each other, one plate has to yield. It is always the more dense, thinner, oceanic plate that subducts beneath the lighter continental plate and is eventually destroyed through melting in the mantle. Such a boundary is known as a consuming boundary. It always produces a trench along the boundary and a certain amount of buckling along the edge of the continental plate, cf. FIG. 19. The intensively studied Japan Trench shows relative horizontal movement of about 7.5 cm per year. The *coseismic displacements* can reach tens of metres horizontally and several metres vertically. For illustration, horizontal displacements resulting from the 1964 tectonic earthquake in Alaska, as calculated by WHITTEN [1970], are given in FIG. 20.

Considerable vertical movements, on the edge of the continental plate, have been reported by TSUBOI [1933] ranging up to several decimetres in the course of a few decades. The two kinds of such movements, preseismic and coseismic, are shown diagrammatically in FIG. 21 (according to RIKITAKE [1976]). A collision of two continental plates is the main (orogenic) mountain building process. Since neither of the plates can underthrust the other because of their buoyancy, the collision results in enormous buckling.

The last type of plate boundary is the *transcurrent boundary*. The relative movements of such plates may be either unobstructed or temporarily arrested by friction along the scar. The arrested motion results in stress accumulation in the zone adjacent to the boundary and is sooner or later released in the form of an earthquake. This is, of course, true even of the consuming boundaries (as seen in

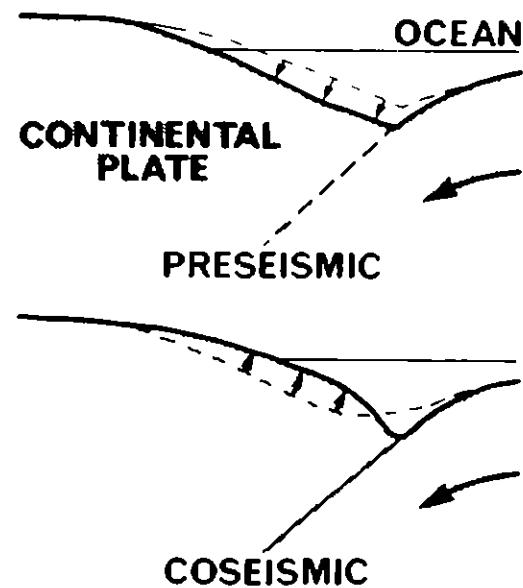


FIG. 8.21. Compression and rebound of continental plate.

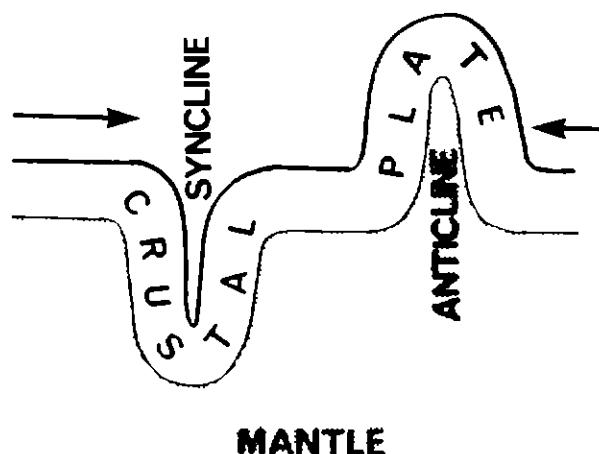


FIG. 8.22. Geosynclines.

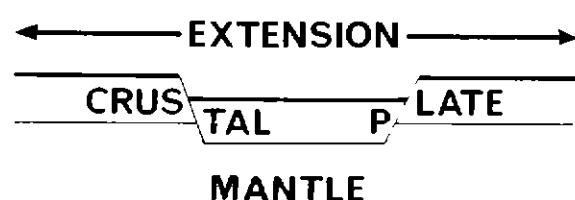


FIG. 8.23. Grabens.



FIG. 8.24. Escarpments.

FIG. 19), and thus most of the world's earthquakes occur in the zone around these boundaries. The stress build-up can be observed on the surface of the earth as strain or as vertical displacements—swells and depressions of the ground. This phenomenon is the basis for the idea of using geodetic methods for earthquake prediction. In the case of the best known example of a transcurrent boundary—the San Andreas fault along the California coast—the observed rate of creep is 3.2 cm per year [SAVAGE AND BURFORD, 1973]. At places, horizontal displacement experienced during the 1905 San Francisco earthquake exceeds 5 m [HAYFORD AND BALDWIN, 1907].

In passing, it should be mentioned that there are other conspicuous manifestations of the plate movements. One of them is the development of *geosynclines*, synclines or anticlines (cf. FIG. 22), as a product of lateral stresses in the crust. Little is known about the ongoing developments, but geological evidence from the past is abundant. Shear stress is also responsible for *faulting*, i.e., tears in the crust. The geometry of the active faults can tell much about the stress pattern in the crust which, in turn, helps to locate the plate boundaries as well as determine the sense of relative motion of the plates (unless faults occur along lines of structural weakness).

Faulting is not, however, confined to regions around plate boundaries. Faults develop even within the plates, presumably also due to the plate movement. It is speculated by some scholars, e.g., MENARD [1973], that the lithospheric plates may be dragged over an uneven interface with the asthenosphere. Such motion would introduce additional shear stresses, which would then result in faulting within the plates. Two notable configurations thus produced are *grabens* (FIG. 23) and *escarpments* (FIG. 24). Again, little is known about the processes presently going on. Menard estimates, however, that the maximum rate of vertical movement induced by the plates sliding over an uneven surface, even coupled with the cooling effect of the mantle, would be smaller than 2 cm per century.

#### 8.4. Man-made and other deformations

Another kind of deformation that takes place in the uppermost layers of the earth's crust is due to *ground compaction*. This deformation manifests itself predominantly as a local or regional subsidence. Apart from loading, which has been dealt with in §8.2, the main cause of ground compaction is the withdrawal of fluids from the ground. Excessive extraction of underground water usually results in sizeable settlements over comparatively large areas.

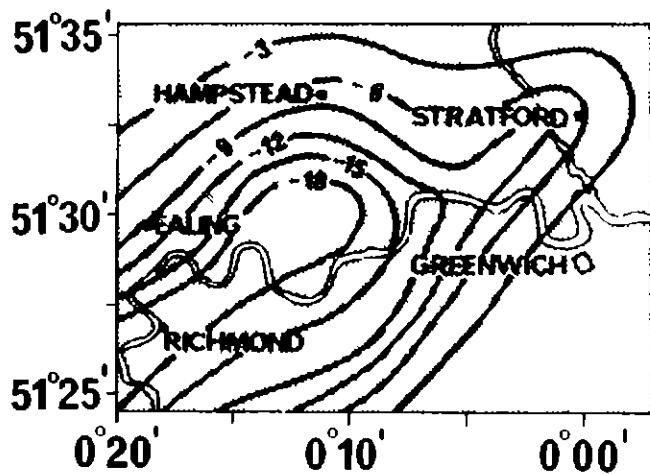


FIG. 8.25. Subsidence of the City of London (U.K.). Contours in centimetres.

By way of illustration, FIG. 25 shows the subsidence of the City of London during the period 1865 to 1931, as determined by WILSON AND GRACE [1942]. The resulting change in vertical positions is by no means negligible and must be taken into account in geodetic works, as well as in various projects.

California provides another example of a regional subsidence due to the extraction of underground water. In the San Joaquin Valley, an area of about 15 000 km<sup>2</sup> has

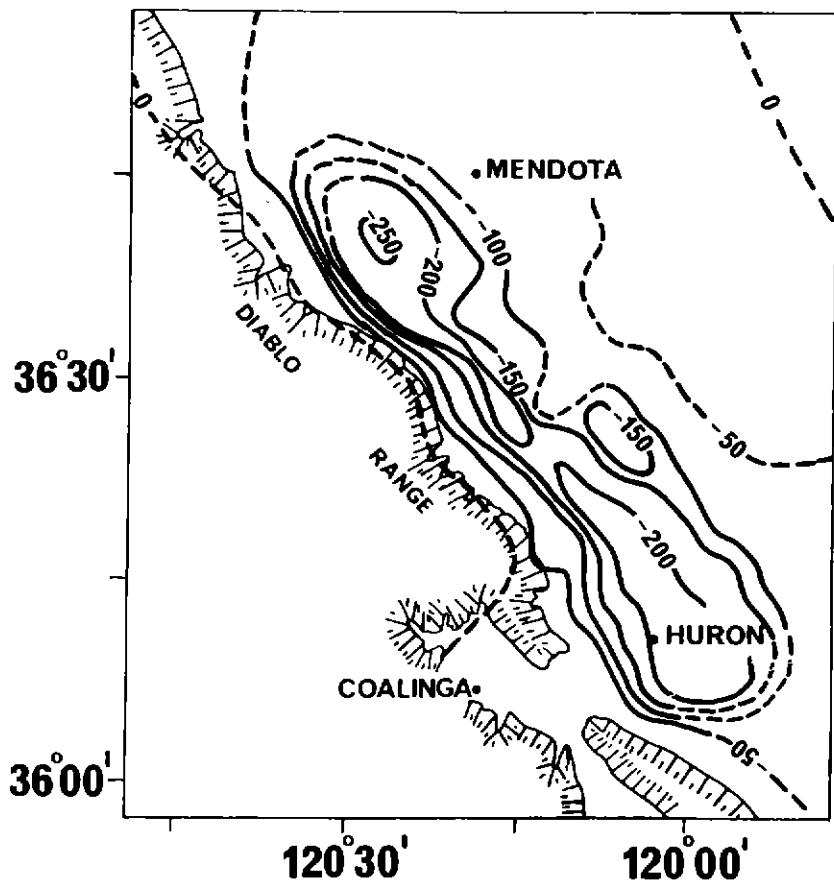


FIG. 8.26. Subsidence of the San Joaquin Valley, California (U.S.A.). Contours in centimetres.

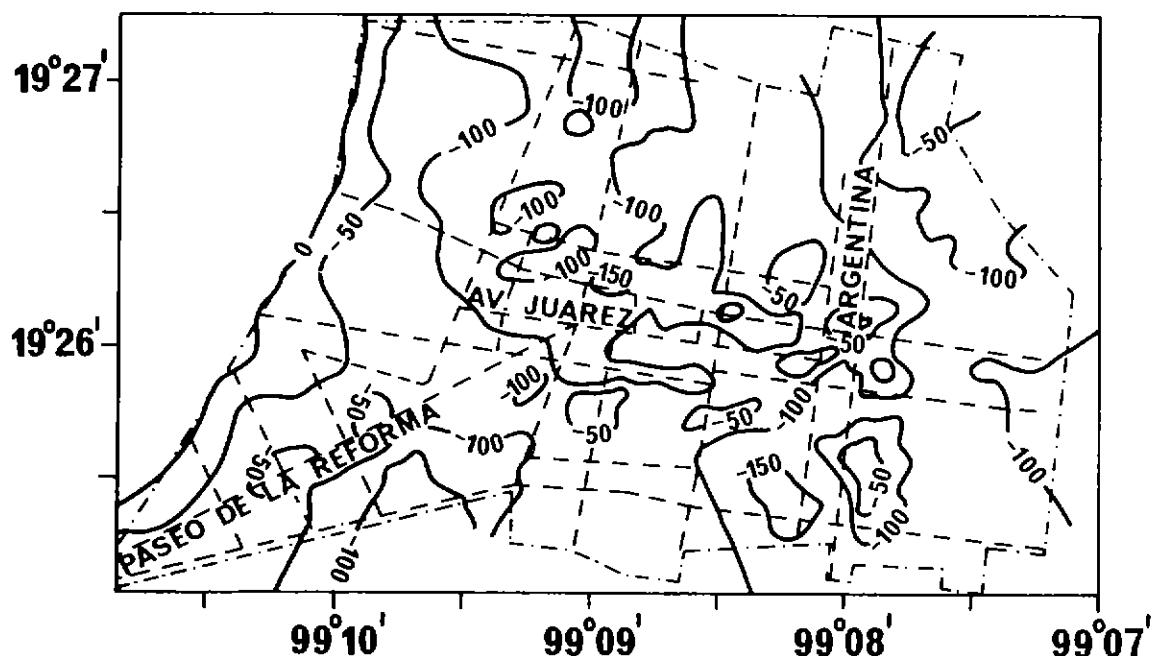


FIG. 8.27. Subsidence of Mexico City (Mexico). Contours in centimetres.

experienced subsidence. The amount of subsidence in the northern part of the Valley during the period 1959 to 1969 is shown in FIG. 26 [HOLDAHL, 1969]—the maximum reaching over 2.5 metres.

Similar magnitudes of settlement have been observed in Mexico City. Subsidence of up to 1.5 m (cf. FIG. 27) occurred during the period 1952 to 1957 [COMISION HIDROLOGICA DE LA CUENCA DEL VALLE DE MÉXICO, 1961]. Since 1891, there has been an overall accumulation of 7.5 m as a result of draining the surrounding lake, and the ever-increasing consumption of underground water. The subsidence still goes on at a high rate.

Ground settlement in the vicinity of active oil and gas fields is also, at least in part, of the same origin. FIG. 28 shows the amount of overall settlement across the Wilmington oil field in Los Angeles which, according to YERKES AND CASTLE [1971], is thought to be due to the oil extraction during the years 1928 to 1962. A surprising feature of this investigation is the large horizontal displacement, of several metres,

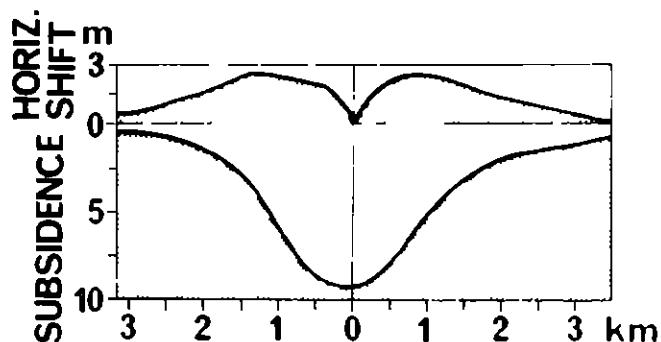


FIG. 8.28. Ground deformation across Wilmington oil field, California (U.S.A.).

associated with the vertical subsidence. This deformation may be partly due to the caving in of emptied underground cavities. Observations have shown that at least some of the subsidence is recoverable by pumping fluid back into the ground [POLAND AND DAVIS, 1969].

*Caving in* of underground cavities, both natural and man-made, is another well-known source of subsidence that can affect relatively wide areas on the surface of the earth. An opposite effect, *ground swelling*, has been observed in connection with underground fluid waste disposal. When the waste is injected into deeper crustal strata, the uplift can significantly affect large regions.

One of the most conspicuous of the phenomena that drastically changes the shape of the earth is earthquakes. The largest earthquakes are related to the tectonic plate movements, as already discussed in §8.3. However, not all earthquakes occur along the plate boundaries. For dramatic effects, *landslides* of various origin can be compared to earthquakes. Because of the large mass involved in the movement, the landslides may even have a large local loading effect. Both earthquakes and landslides cause a local change in the gravity field.

To bring the list of significant deformations to an end, it should be mentioned that, occasionally, regional *anomalous uplifts or subsidences* are discovered that have no immediately obvious origin. One such uplift occurred on the border between the American states of Mississippi and Alabama. Geodetic data, collected since 1900, point to a velocity of the uplift of over 50 cm per century [HOLDAHL AND MORRISON, 1974]. Another such region was discovered by FROST AND LILLY [1966] in the Lac St. Jean area of Québec, Canada. Subsequent analyses by GALE [1970] and VANÍČEK AND HAMILTON [1972] confirmed their findings. The result of the latter analysis is given in FIG. 29 in terms of vertical velocities. There is no universally accepted geophysical explanation for these anomalous movements. The

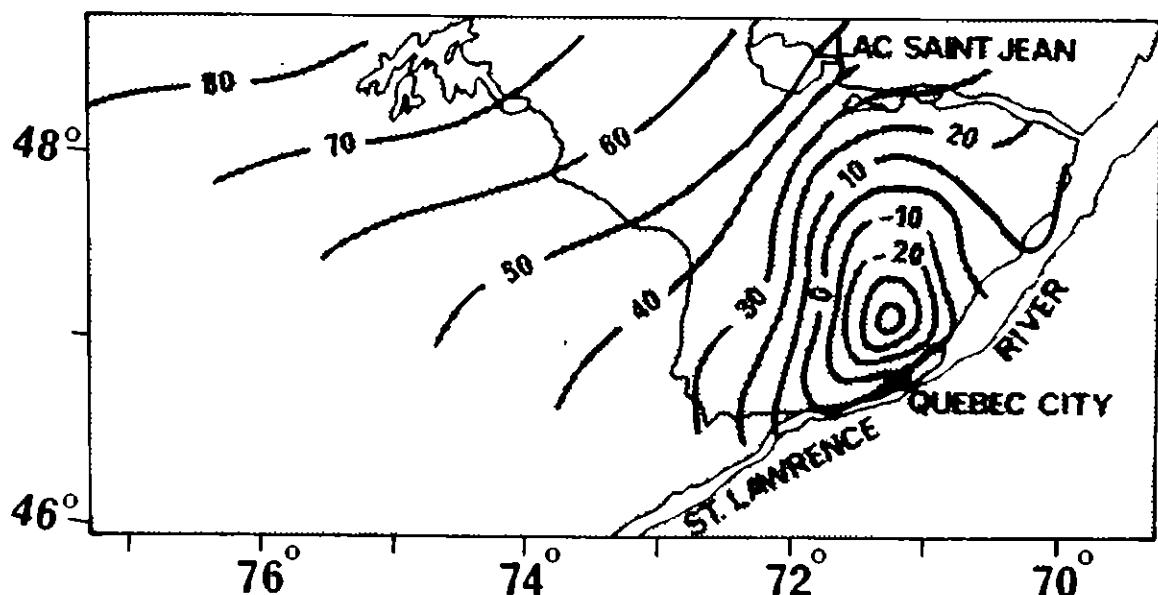


FIG. 8.29. Subsidence in Québec (Canada). Contours in millimetres per year.

fact that they are not readily explainable makes these movements unpredictable and thus particularly hazardous to geodetic operations. The possibility cannot be ruled out that at least some of these movements are of tectonic origin.

There are scores of phenomena of minor significance to geodesy. Of at least theoretical interest among these are various vibrations and tremors of different provenience. The best known are *seismic waves* — an effect of earthquakes that can be felt the world over. On the sea, seismic and other events generate long wavelength waves known as *tsunamis* (see FIG. 31). Man-made and other *tremors* are high-frequency occurrences peculiar to the locality. These are merely a nuisance in numerous very precise geodetic measurements where they appear as high frequency noise.

Also of an oscillating nature are the deformations of the earth (and its gravity field) induced by *polar motion*. While the polar motion itself does not change the shape of the earth, its effects on the earth's shape through the changes of centrifugal potential (cf. §6.3) are real. The effects are very small and will be dealt with in §25.3. *Free oscillations* of the earth are the vibrations of the whole earth's body. They can be induced by earthquakes or other shocks. FIG. 30 shows a spectrum of free oscillations as observed through gravity variations [NAKAGAWA ET AL., 1968]. Notice that the fundamental mode, with period around 54 minutes, is particularly distinguishable.

All of the deformations described so far influence the earth's gravity field and thus can be observed through gravity variations, as will be seen in Chapter 26. The temporal variations in the gravity field lead us to mention two phenomena that affect gravity directly. The first is *gravitational waves* emitted by celestial bodies outside our solar system [MISNER ET AL., 1973]. These waves are so weak that they are detectable only with specially designed instruments. Another phenomenon is the secular *change of the gravitational constant G* (cf. §6.1) predicted by some cosmological theories [WILL, 1971]. This change, if it exists, is again very slow and extremely small. Both of these effects are of academic interest only.

We shall return to most of the above described phenomena in Part VI, with a more systematic discussion of their effect on geodetic work as well as the role geodesy plays in their detection.

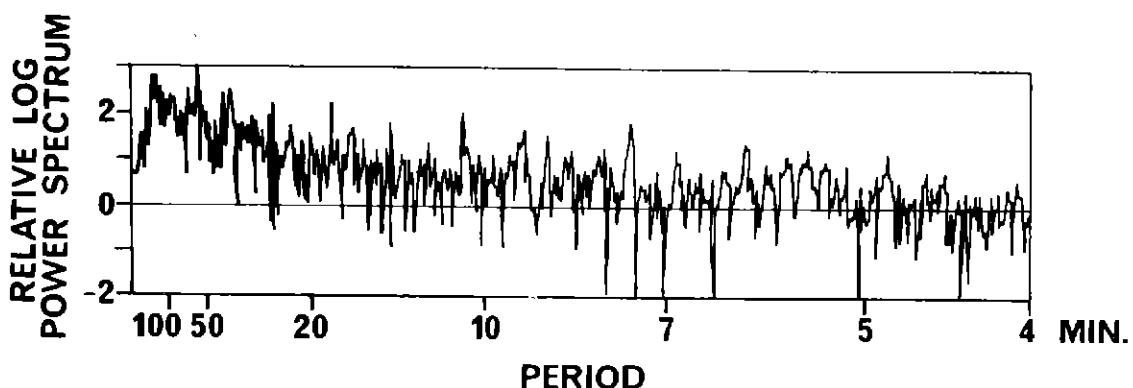


FIG. 8.30. Free oscillations of the earth.

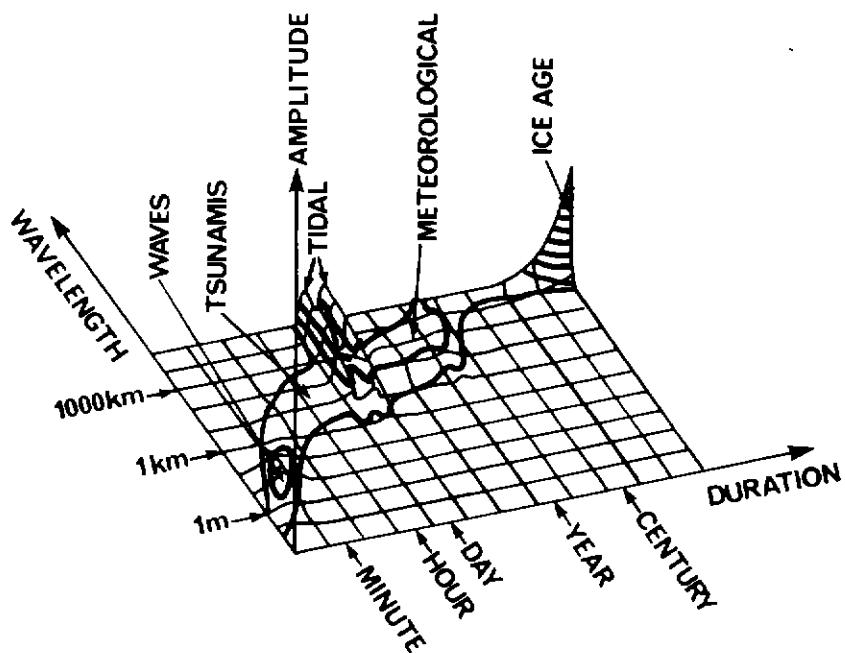


FIG. 8.31. Sea level temporal variations.

For the sake of completeness however, a few words will be said about the *time variations of the sea level*, an understanding of which is needed for vertical positioning (§19.1). A two-dimensional, general spectrum showing the frequencies, amplitudes, and regional extent of the diverse kinds of sea level variations is given in FIG. 31, according to STOMMEL [1963]. Note that this spectrum includes the tidal terms and the variation due to ice melting at the end of an ice age. It also includes different kinds of short periodic and short-lived phenomena caused by meteorological and other effects. These have little impact on geodesy and will not be discussed further.

On the other hand, the long-periodic, and particularly the secular, changes in sea level are important in geodesy. The *secular changes* directly affect the definition of the geoid and thus indirectly other quantities. The reverse is, of course, also true, and the discussion pertaining to FIG. 8 above should always be borne in mind when secular changes to sea level are discussed. The global, mean secular (*eustatic*) sea level variations in the period 1860 to 1960, as estimated by FAIRBRIDGE AND KREBS [1962], are shown in FIG. 32. The most probable contributors to the eustatic sea level changes are the melting of Antarctica's, as well as other, permanent ice sheets on the

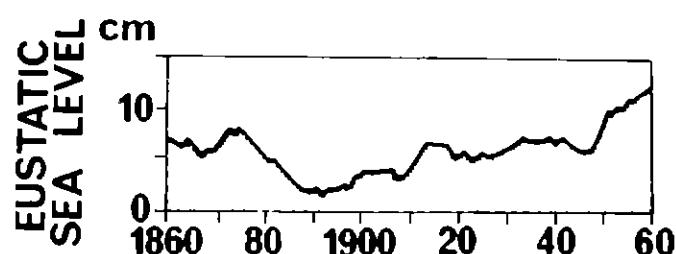


FIG. 8.32. Global mean (eustatic) sea level variations.

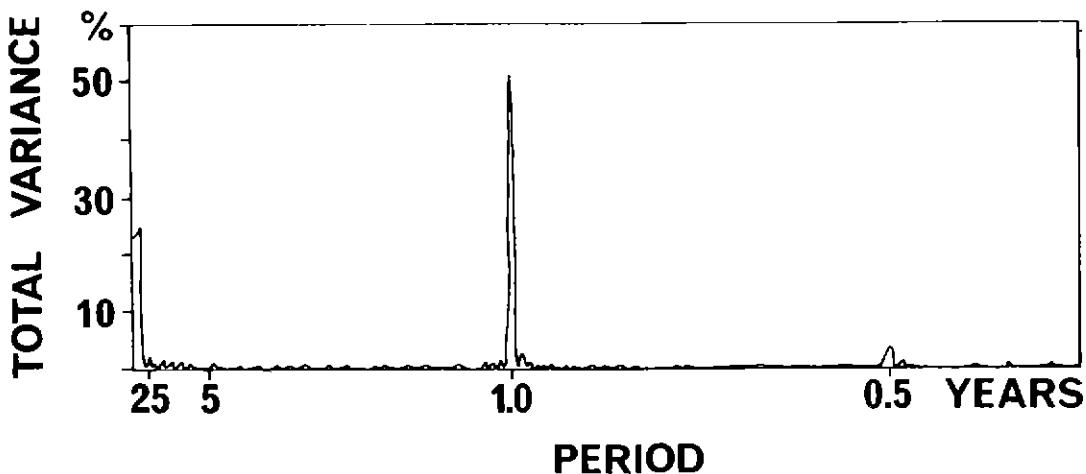


FIG. 8.33. Spectrum of monthly averages of sea level in Baltimore, Maryland (U.S.A.).

surface of the earth, and the continuing adjustment of the lithosphere-asthenosphere system to the load of water freed after the last glacial melt.

The other long-periodic variation of interest is the *annual variation*. It amounts to several decimetres for some places and is known to originate primarily from annual variations in temperature, pressure, wind direction, and wind magnitude. Some records from tide gauges located close to river estuaries display a prominent

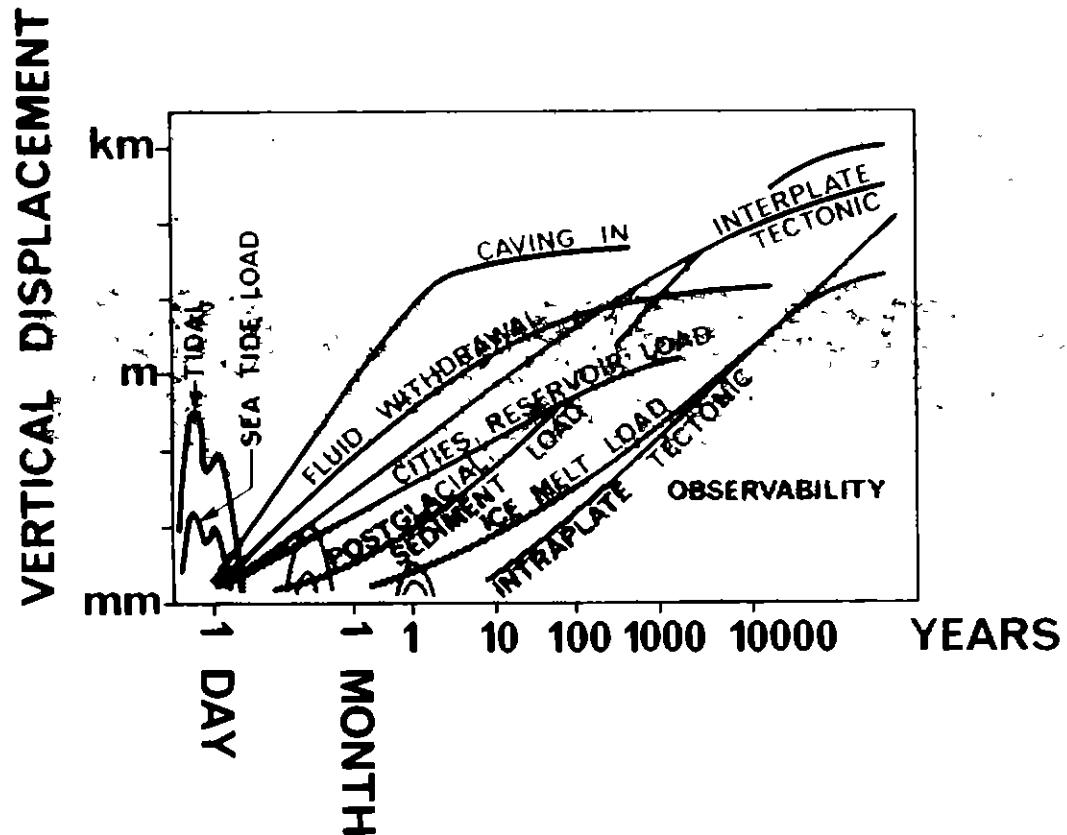


FIG. 8.34. Characteristic vertical displacements due to the most conspicuous phenomena.

semi-annual variation thought to be due to the precipitation cycle. Other frequencies related to different meteorological effects can also be seen on the spectrum shown in FIG. 33 [QURAISHEE AND VANÍČEK, 1970].

To finish this chapter, it is illustrative to compare the different kinds of maximum vertical displacements experienced on the surface of the earth. The most conspicuous displacements, which can be at least roughly quantified at present, are shown in FIG. 34.

## CHAPTER 9

### EARTH AND ITS ATMOSPHERE

Most geodetic measurements—terrestrial and extra-terrestrial—made on the surface of the earth are influenced to a varying degree by the presence of the atmosphere. Consequently, it is necessary to have at least an elementary understanding of some of the processes that occur within the atmosphere.

This chapter contains four sections. The first describes the atmosphere and its basic physical characteristics. The second section deals with the propagation of electromagnetic waves through the atmosphere and, in the absence of any better place to treat it, with the propagation of sound waves through water. It presents the physical laws upon which the formulae for refraction used in geodesy are developed. The third section gives the fundamentals of the dynamics of the atmosphere. The authors believe that some understanding of this topic would be useful to anyone intending to take a serious interest in problems connected with the sea level and with the movements of the earth's crust. In the last section, a more contemporary topic, that of the gravity field generated by the atmosphere and the effect on the earth's gravity field, is introduced

#### 9.1. Some physical properties of the atmosphere

The term *atmosphere* is the accepted name for the air masses enveloping the earth. It is known that the air density decreases with increasing altitude; at an altitude of 600 km to 1000 km, it generally can be regarded as having vanished almost completely. As explained in Chapter 6, the earth's gravitational attraction is weaker with altitude. With increasing height, the influence of the earth's magnetic field grows until it completely dominates the behaviour of air particles in the space adjacent to what is usually considered to be the upper atmospheric boundary. The reasons for this fact will become clear later.

The atmosphere is composed primarily of three gases: nitrogen (approx. 78%), oxygen (approx. 21%), and argon (approx. 1%) [PETTERSEN, 1969]. Also present are traces of other gases and particles, with carbon dioxide and ozone being the most important. In the atmosphere as a whole, most of the atoms are electrically inert. Some, however, are ionized through exposure to various kinds of radiations coming from space. Thus the concentration of these ions generally increases toward the outer reaches.

Let us first have a look at the most conspicuous physical parameter of the atmosphere—the *air temperature*. As we all know, the atmospheric temperature on the surface of the earth varies from point to point. It also varies with time, being influenced by two main cycles—the seasonal (annual), due to the earth's motion around the sun, and the diurnal, due to the spin of the earth. The minimum temperature observed on the earth's surface during the past century was  $-89.2^{\circ}\text{C}$  ( $-128.6^{\circ}\text{F}$ ) recorded on July 21, 1983 at Vostok, Antarctica. The maximum value of  $58.0^{\circ}\text{C}$  ( $136.4^{\circ}\text{F}$ ) was recorded on September 13, 1922 at el-Azizia, Libya [THE WORLD ALMANAC AND BOOK OF FACTS 1984, 1983].

The atmospheric temperature also varies considerably in the vertical direction. FIG. 1 shows the global average distribution of temperature according to PETTERSEN [1969]. Temperature is the parameter delineating the conventionally accepted division of the atmosphere into specific layers. The lowermost layer, taken as being 8 to 17 km thick, is called the *troposphere* and is bounded from above by the *tropopause*. Most of the known meteorological phenomena, like winds, clouds, fog, etc., occur in this layer. The air here is unsettled, and its upper boundary varies considerably with latitude and season. The next stratum, extending from the tropopause to about 50 km, is known as the *stratosphere* and is bounded from above by the *stratopause*. The air distribution in the stratosphere is very stable, and the air is dry. The lowest 30 km of the atmosphere contain about 99% of all the air molecules. The *mesosphere*, the name of the layer above the stratosphere, is characterized by the turbulent motion of the air. It stretches to an altitude of 80 km, where the *mesopause* is located. Further up one encounters the *ionosphere* with its characteristically higher

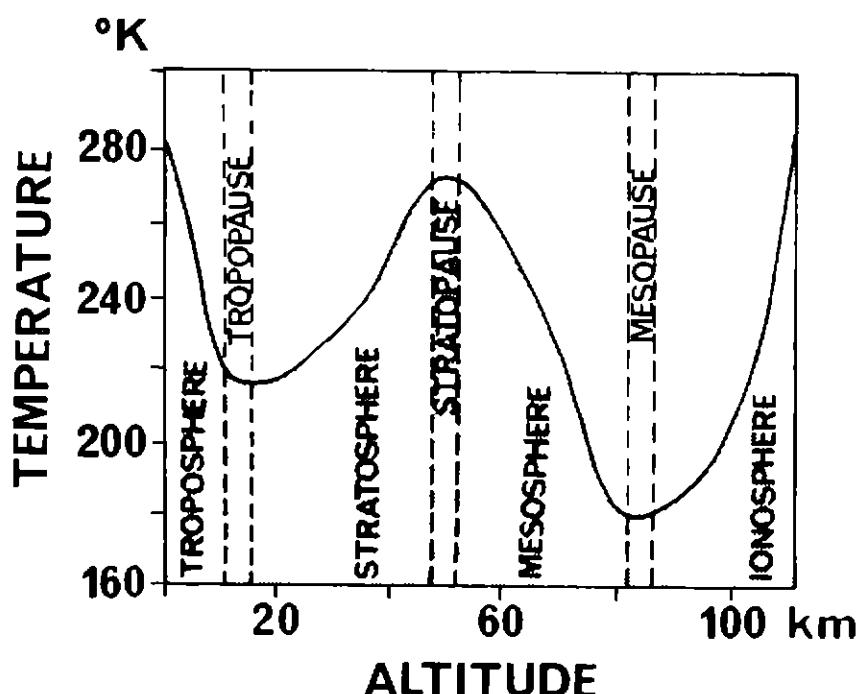


FIG. 9.1. Vertical distribution of temperature in the atmosphere.

concentration of ionized gases. From an altitude of 400 km to the limits of the earth's magnetic field is the layer known as the *magnetosphere*. Note that different nomenclatures for the division of the atmosphere are used by other authors.

For solving various problems, particularly for the troposphere, the *vertical temperature gradient* is needed. Its global, average, tropospheric value is about  $0.0065^{\circ}\text{C}$  per metre. Locally, however, the gradient varies considerably with location and time.

The second physical parameter that needs be studied is the *air density* of the atmosphere. Like temperature, density varies not only from place to place but also with time. Most conspicuously, it decreases rapidly with altitude. Various institutions have proposed stationary models for the global density distribution: COSPAR (International Reference Atmosphere)—CIRA [COSPAR, 1965], U.S. National Advisory Committee for Aeronautics—NACA [DIEHL, 1948], Smithsonian Astrophysical Observatory—SAO [LIST, 1958], U.S. Air Research and Development Command—ARDC [MINZNER ET AL., 1959], to name a few. These models normally include other parameters like temperature and pressure distribution. FIG. 2 illustrates the density distribution used in the NACA model. Note that even the densest layer of the atmosphere is only about 0.12% of the density of water and 0.04% of the density of surface rock (cf. §8.2).

The air density can be measured by (barometric) *air pressure*, which is merely the hydrostatic pressure, or weight of a column of air on a unit area. It is easy to see that the weight  $\Delta w$  of a column of air,  $\Delta h$  in height, is given by

$$\Delta w = \text{mean}(\sigma_a g) \Delta h, \quad (9.1)$$

where  $\sigma_a$  is the density of air, and the mean is taken along  $\Delta h$  in this column. Integrating the weight along a plumb line from a given height  $h$  to the limits of the atmosphere (or, e.g., 40 km with an accuracy of 99.7%), we obtain the pressure at

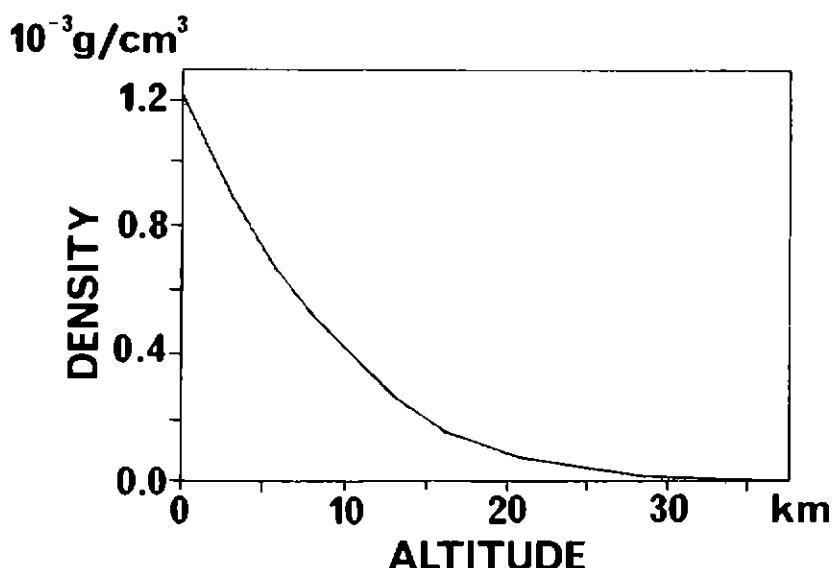


FIG. 9.2. Vertical distribution of density in the atmosphere.

height  $h$ :

$$p \doteq \int_h^{40 \text{ km}} \sigma_a g dh. \quad (9.2)$$

Since the decrease of pressure with altitude at any given time is quite regular, pressure can also be used as an approximate measure of height as long as care is taken of the fact that pressure changes with time in response to variations in the density distribution. The barometric determination of heights (see §19.4) is based on this principle.

Barometric pressure is usually measured in bars (1 bar equals  $10^{-5} \text{ N cm}^{-2}$ , or 100 kPa, and is equivalent to the pressure exerted by 0.75006 m of mercury). Normal pressure on the earth's surface is around 1 bar, or 1000 millibars. Surfaces of equal pressure are called *isobaric surfaces*. Lines joining points of equal pressure (for this purpose usually reduced to sea level) on the earth's surface are known as *isobars*. Because of the temporal variations of density, both isobaric surfaces and isobars vary with time. As an example, FIG. 3 gives the average global pressure for the month of January, according to the MEDALLION WORLD ATLAS [1973].

Temperature  $T$  (absolute in degrees K) and pressure  $p$  are related through the *equation of state* of an ideal gas that reads [MENZEL, 1955]:

$$p = \sigma_a \frac{R}{m} T, \quad (9.3)$$

where  $R$  is the universal gas constant equal to  $8.31696 \times 10^7 \text{ erg mol}^{-1} \text{ K}^{-1}$ , and  $m$  is the molecular weight of the air. This equation is, strictly speaking, valid only for dry air, for which the ratio  $R/m$  equals  $2.8704 \times 10^6 \text{ cm}^2 \text{ s}^{-2} \text{ K}^{-1}$  [LIST, 1958].

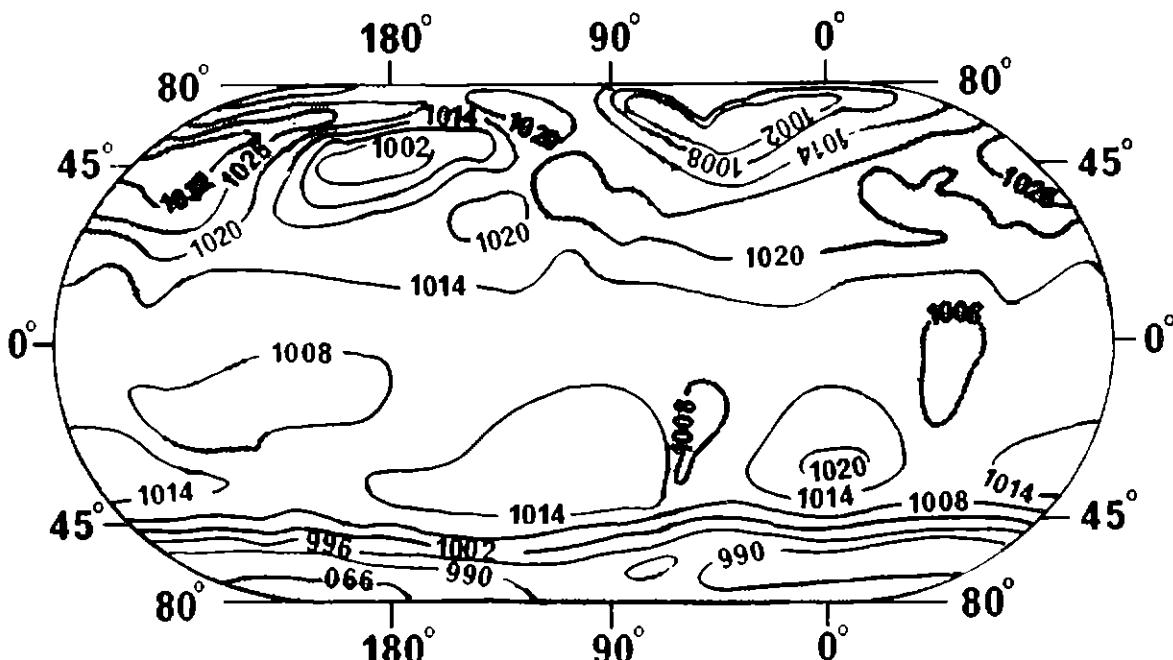


FIG. 9.3. Average global pressure for the month of January. Contours in millibars.

In the troposphere, however, the air is seldom absolutely dry. There is usually a moisture content in the air which, under certain conditions, shows as clouds, fog, mist, etc. For moist air, the equation of state holds only approximately; more appropriate is the following form:

$$p \doteq \sigma_a \frac{T}{1 - 0.37803 e/p} \cdot \frac{R}{m}, \quad (9.4)$$

where the humidity is expressed in terms of the partial pressure  $e$  of the water vapour content. The amount of *air humidity* can also be calculated from wet and dry bulb temperatures or obtained from a hygrometer.

## 9.2. Wave propagation through the atmosphere and water

Electromagnetic waves, employed in making geometrical measurements in the atmosphere, vary from radio-frequency waves through microwaves to visible light, the corresponding frequency range being between  $10^4$  and  $10^{15}$  Hz. In some respects, sound waves (with frequencies between 10 and  $10^4$  Hz), used for measuring distances in water, behave in a fashion similar to that of the electromagnetic waves. Thus, in this section, the concepts of propagation of both kinds of waves are treated.

Let us focus our attention first on the propagation of electromagnetic waves through the atmosphere. Depending on the frequency  $f$ , three distinctly different propagation characteristics are recognized, as shown on FIG. 4. While waves of all frequencies propagate in the *direct wave* mode, where intervisibility between the transmitter and receiver is required, the lower frequency (long wavelength) waves are also capable of travelling in the *ground wave* mode. For direction determination, the usefulness of the direct wave is obvious; nevertheless, the ground wave can be used as well if the direction is sought in terms of the strongest signal. For determining distances, one can use either the direct or the ground waves. Thus the determination of either a direction or a distance is not limited by the intervisibility of the two end points.

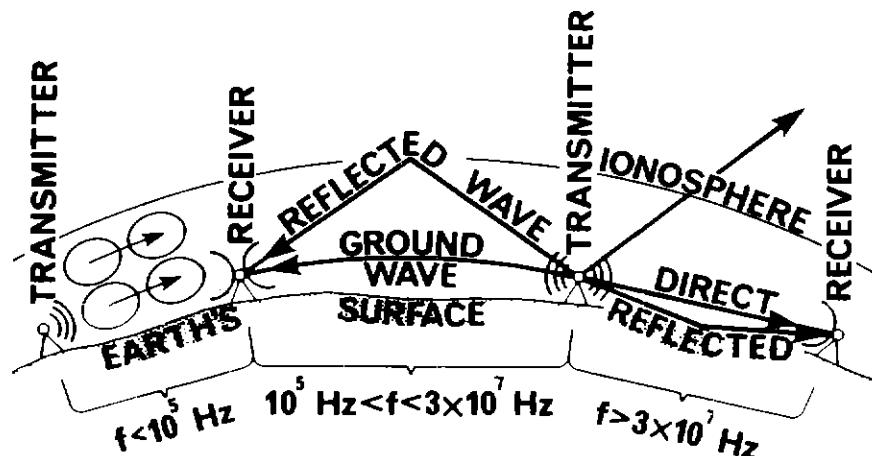


FIG. 9.4. Propagation of electromagnetic waves through the atmosphere.

The *reflected waves* (cf. FIG. 4) are a nuisance in practically all geodetic measurements [BURNSIDE, 1971]; it has been shown recently, however; that the wave reflected off the ionosphere can be useful for certain purposes [WELLS, 1979]. The lower frequency waves are reflected by the lowest stratum of the ionosphere so that they cannot penetrate into outer space. On the other hand, the propagation of the high frequency (over  $3 \times 10^7$  Hz) waves is not curbed by the ionosphere. All the waves are 'reflected' from the ground to a certain extent, although for the waves of very low frequencies, one really cannot speak about reflection. Perhaps 'rolling' might be a better term. Excepting the very low frequencies, the level of reflection increases with the wavelength, the conductivity of the ground, and the angle of incidence; e.g., it is almost perfect for frequencies below  $10^{10}$  Hz coming at a right angle to the water surface [HILL, 1966].

As stated in §7.3, all the modern measurements of distance are based on timing the propagation of an electromagnetic wave between two points. Unfortunately though, the velocity of propagation changes with the propagation medium. The highest velocity,  $c$ , estimated now to be  $299\,792\,458 \pm 12$  ms<sup>-1</sup> [TERRIEN, 1974], is attained in a vacuum. The direction of the wave, used in direction measurements, also changes with the medium; even the direct wave is *refracted* (bent) by the medium. The refraction of the direct wave and the velocity changes are very closely related and are usually treated simultaneously under the term *refraction*. By the refraction of the ground wave we mean only the change in velocity; it clearly does not make sense to speak about directional refraction of the ground wave.

At this stage it is expedient to introduce the *index of refraction*,  $n$ , defined as

$$n = c/v. \quad (9.5)$$

Since the actual velocity  $v$  is always smaller than  $c$ ,  $n$  is always slightly larger than 1; it depends strongly on the density,  $\sigma$ , of the medium (i.e., air in this case) and weakly on the wavelength of the propagating wave. As will be seen later, in the ionosphere the refractive index for modulated waves is smaller than 1, because their velocity is perceived somewhat differently.

The relation between  $n$  and  $\sigma$  is given by the *Lorenz–Lorentz equation* [MENZEL, 1955]:

$$\frac{n^2 - 1}{n^2 + 2} = \sigma r, \quad (9.6)$$

where  $r$  is called the *specific refractivity* of the medium, and it is nearly constant for the usual range of wavelengths. From experiments conducted in the troposphere,  $n$  is expected to have a value between 1 and 1.0003 [HOTINE, 1969]. For this range of values, to a high degree of accuracy, the ratio in (6) reduces to  $\frac{2}{3}(n - 1)$  so that for the tropospheric density  $\sigma_a$  one obtains

$n - 1 \doteq \frac{3}{2} \sigma_a r.$

(9.7)

Because the actual velocity  $v$  is lower than  $c$ , the travel time of the electromagnetic wave, used to measure a distance between two points in the atmosphere, is longer

than it would be in a vacuum. Thus the distance appears more extended than it actually is; the closer one is to the ground (i.e., the denser the air), the longer the distance appears. The matter is complicated by the fact that the density  $\sigma_a$  varies somewhat irregularly and, as a consequence, so does the refractive index.

The general law governing the geometry of the (direct) wave path has been formulated by Fermat; it is a variation of the more general Hamiltonian principle of minimum energy [CONDON AND ODISHAW, 1967]. *Fermat's principle* ensures that through a medium the travel time of the wave between any two given points is the minimum. Considering an infinitesimally short increment of the path  $dS$ , the instantaneous velocity of propagation is given as

$$v = \frac{dS}{d\tau}. \quad (9.8)$$

Substitution of  $c/n$  for  $v$  from (5) and integration over any path  $\mathcal{C}$  between the two end points  $P_1, P_2$ , results in the following expression for the travel time  $\tau_2 - \tau_1$ :

$$(\tau_2 - \tau_1)_{\mathcal{C}} = \frac{1}{c} \int_{\mathcal{C}} n dS. \quad (9.9)$$

The direct electromagnetic wave follows the path  $\tilde{\mathcal{C}}$ , which minimizes the above expression—see FIG. 5. The product  $c(\tau_2 - \tau_1)_{\tilde{\mathcal{C}}}$  is the minimum path length, called the *eikonal* in optics.

The spatial curve (the path  $\vec{r}(S)$ ) is obtained by solving the variational problem described by (9). Adopting an arbitrary Cartesian coordinate system and expressing the refractive index as a function of position (i.e.,  $n(\vec{r})$ ), one arrives at the following differential equation for the curve [MENZEL, 1955]:

$$\frac{d}{dS} \left( n \frac{d\vec{r}}{dS} \right) - \text{grad } n = 0.$$

(9.10)

The resulting curve,  $\vec{r}(S)$  (see §3.3) obviously depends on  $n$  and possesses both curvature (bend) as well as torsion (twist); its length, the eikonal, is readily obtained from (9).

It is the deviation of the shortest path from the straight line connecting  $P_1$  with  $P_2$  (i.e., the proper refraction) that is of the utmost interest in direction determination. When attempting to solve eqn. (10) one realizes that, aside from having to specify

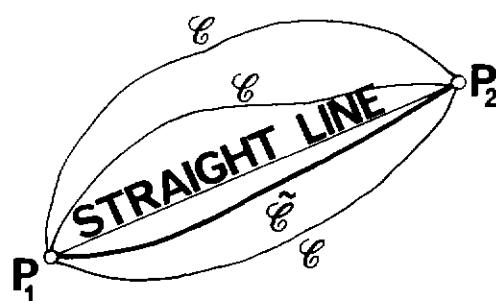


FIG. 9.5. Fermat's principle.

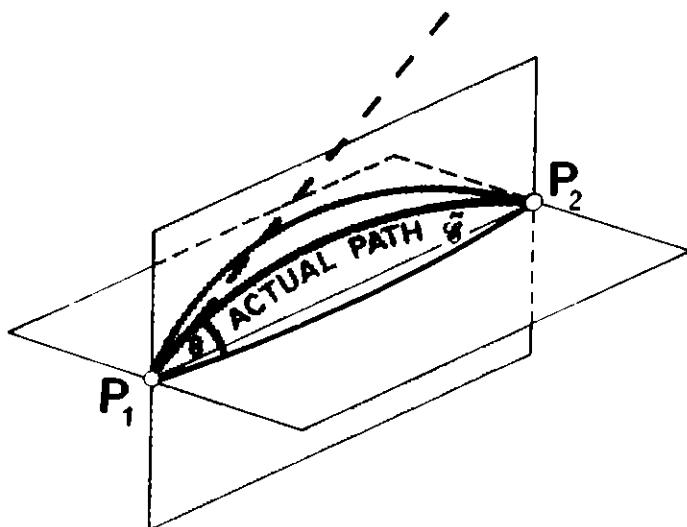


FIG. 9.6. Vertical and horizontal refraction.

the wavelength of the electromagnetic wave, a complete knowledge of the air density  $\sigma_a$  along the path is also required. Since this is not directly measurable, the density is calculated from (4) using the observed temperature, pressure, and partial pressure of water vapour. In practice, it is not easy to obtain all this information, thus, simplifications have to be made.

To begin with, let us consider the vertical and horizontal planes containing the two end points  $P_1$  and  $P_2$  of the path (FIG. 6). We have seen in §9.1 that the air density varies predominately in the vertical sense. Hence, the path is more refracted in the vertical sense, i.e., the projection of the path onto the vertical plane is more refracted than the projection onto the horizontal plane. In other words, the angle between the straight line and the actual path  $\vec{C}$ —the *refractive angle*—is larger in the vertical sense. Observations indeed indicate that directional refraction in the vertical sense, called the *vertical refraction*, is generally at least one order of magnitude larger than the *horizontal refraction*. Magnitudes of the refractive angle  $\theta$  greater than one minute of arc have been recorded [PELIKÁN, 1967] but may conceivably be much larger. More typically, the refractive angles are of the order of  $10''$ .

Clearly, in the first approximation, the air density can be regarded as horizontally stratified. If this stratified atmospheric model along the path is accepted, the curvature of the path in the vertical plane can be derived as follows. *Snell's law* states that the passage through a boundary between two strata (cf. FIG. 7) with

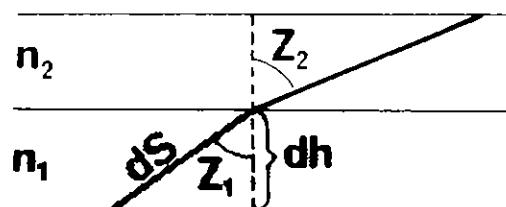


FIG. 9.7. Refraction at a boundary of two layers.

refractive indices  $n_1, n_2$  is governed by the following equation [DRUDE, 1959]:

$$n_1 \sin Z_1 = n_2 \sin Z_2. \quad (9.11)$$

Taking the thickness  $dh$  of the stratum to be infinitesimally small and the refractive index to be changing continuously, i.e.,  $n_2 - n_1 = dn$ , one obtains

$$Z_2 - Z_1 = dZ \doteq - \frac{dn}{n} \tan Z. \quad (9.12)$$

On the other hand, the radius of curvature  $R$  (not to be confused with the mean radius of curvature of the earth) of a circle osculating to the path at  $P$  is given as

$$\frac{1}{R} = \frac{dZ}{dS}. \quad (9.13)$$

Realizing that  $dS \cos Z = dh$ , one finally obtains

$$\boxed{\frac{1}{R} \doteq - \frac{1}{n} \frac{dn}{dh} \sin Z.} \quad (9.14)$$

Note that the vertical refraction, like the retardation of the propagation velocity, is most pronounced when the wave travels in a horizontal direction ( $Z \rightarrow \frac{1}{2}\pi$ ) but, unlike the delay, disappears altogether (in the first approximation) when the wave propagates in a vertical direction ( $Z \rightarrow 0$ ).

In many practical cases, it is sufficient to regard the path of the direct electromagnetic wave between two points close to the ground as circular. This means that the curvature is assumed constant between the two points, and the situation depicted in FIG. 8 occurs.

Evidently, the refraction phenomenon is particularly pronounced in the lower layers of the atmosphere. This is why there it is often referred to as the *tropospheric refraction*. Above the 30 km level, refraction is so weak that it can be neglected altogether. On the other hand, when the electromagnetic waves reach altitudes where

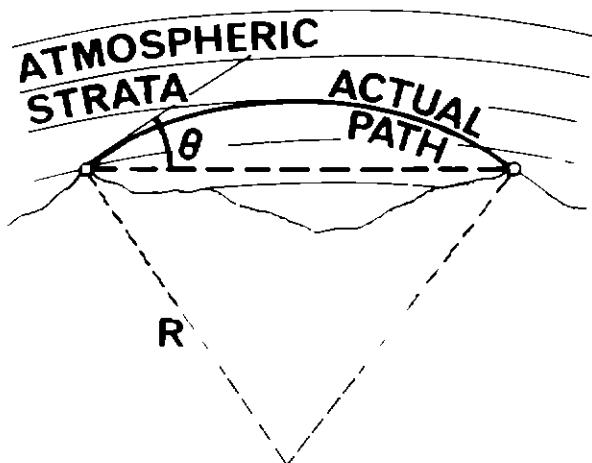


FIG. 9.8. Simplified model of vertical refraction.

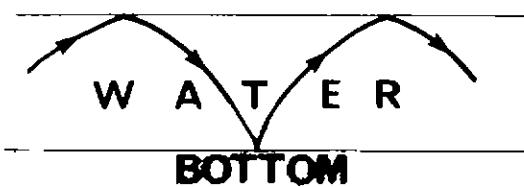


FIG. 9.9. Propagation of sound when velocity decreases with depth.

the air is richer with ions—particularly the ionosphere—another phenomenon takes place. If the wave of a frequency higher than  $3 \times 10^7$  Hz passes through the atmosphere, its propagation velocity is affected by ions in such a way that, to a certain extent, it becomes frequency dependent. This phenomenon is referred to as *dispersion*. The dispersive index of refraction is given by [WEIFFENBACH, 1967]

$$n = \sqrt{1 - \frac{f_N^2}{\alpha f^2}}, \quad (9.15)$$

where  $f_N$  is the electron plasma resonance frequency that depends on electron density and changes with time and position,  $f$  is the electromagnetic wave frequency, and  $\alpha$  is a function of the direction of propagation and the inclination and intensity of the earth's magnetic field. Because the dispersion takes place predominantly in the ionosphere, it is often called the *ionospheric refraction*.

It is interesting to note that the ionospheric index of refraction for modulated waves is, as already stated, smaller than 1. This would indicate a propagation velocity higher than  $c$ —a clear violation of one of the basic postulates of physics. The explanation is that, with a modulated wave, it is the modulation envelope that is the carrier of information. Since the envelope is the result of putting a group of simple waves together, the velocity of the envelope propagation depends on the modulated wave velocity and also on the relative retardation of the modulating waves. It is this modulation envelope velocity, or *group velocity*, that is higher than  $c$  [LE MÉHAUTÉ, 1976].

We have already seen that water is a very efficient reflector for electromagnetic waves. How does it behave as a medium for their propagation? Not very well! Being much more conductive than the air, it tends to attenuate the propagation very



FIG. 9.10. Propagation of sound when velocity increases with depth.

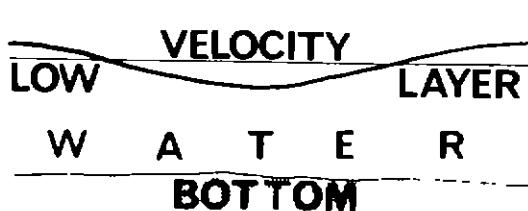


FIG. 9.11 Propagation of sound through a low velocity layer.

rapidly. The higher its frequency, the more the wave is attenuated, and there is little possibility of significant penetration into the water except for frequencies comparable to those of sound (audio frequencies). Also, unlike in the air, the propagation velocity depends strongly on frequency  $\omega$ ; namely [HILL, 1966],

$$v = \sqrt{2\omega/(\lambda\mu)}, \quad (9.16)$$

where  $\lambda$  (not to be confused with longitude) is conductivity and  $\mu$  permeability. For the audio frequency of 1 Hz, the velocity is almost the same as that of sound.

*Sound* propagates somewhat better through water. The propagation velocity is around 1550 m/s and varies significantly with water temperature; less appreciable variations are due to pressure and salinity changes. There is no known theoretical relation between the velocity and the other three parameters. Empirical formulae are used instead, as will be seen in §19.4.

The attenuation, due to the dispersion and absorption (conversion to heat) of energy, is lower in the case of a sound wave than in the case of electromagnetic waves. The main source of attenuation is scatter on an uneven bottom and, to a lesser extent, on an uneven surface. Scatter on air bubbles may also be considerable.

The laws of reflection and refraction for sound in water are analogous to those for electromagnetic waves. Thus, reflections of sound from both the bottom and the surface exist. Propagation modes depicted in FIGS. 9 and 10 then occur. Under special circumstances, when due to a particular water mixing a layer of low propagation velocity deeper below the surface appears, sound propagates within this layer avoiding attenuation through scattering (FIG. 11). Then it may travel for hundreds or thousands of kilometers, as reported by various oceanographers.

### 9.3. Temporal variations of the atmosphere

By temporal variations of the atmosphere, also called *atmospheric dynamics*, we mean the movement of air particles, or temporal variations in air density; these are just two different ways of describing the same phenomenon. The density variations are accompanied by changes in temperature and pressure, and are displayed through winds. The whole field of atmospheric dynamics is, of course, quite complex. Since it is not directly related to geodesy, any deeper discussion of this subject here would be

out of place. The interested reader can pursue this topic in MALONE [1951] or other publications quoted herein. Following is only a very cursory sketch of some of the phenomena pertinent to geodetic work.

The global motion of the atmosphere is approximately described by four equations. First there is the *hydrodynamic equation of motion* [MENZEL, 1955]:

$$\dot{\vec{v}} = -2\vec{\omega} \times \vec{v} - \frac{1}{\sigma_a} \nabla p - \vec{g} + \vec{F}, \quad (9.17)$$

where  $\vec{v}$  is the velocity of the air particles,  $\vec{F}$  is an acceleration that describes the internal stresses, eddy stresses, and viscosity of the atmosphere, and the rest of the symbols have the same meaning as in the previous chapters ( $\vec{\omega}$  being the earth's angular velocity vector, cf. §5.3). The first two terms (on the right-hand side) dominate the motion. The first is known as the *Coriolis acceleration*, and the second is the *pressure gradient acceleration*.

The second equation is called the *equation of continuity*, and its role is to ensure the conservation of atmospheric mass during the motion. It reads:

$$\frac{\partial p}{\partial t} + \nabla \cdot \sigma_a \vec{v} = 0. \quad (9.18)$$

Third is the *first law of thermodynamics* [MENZEL, 1955]

$$dq = C_p dT - \frac{1}{\sigma_a} dp, \quad (9.19)$$

where  $dq$  is the actual specific heat added to the atmosphere, and  $C_p$  is the specific heat of the air at constant pressure  $p$ . Fourth is the already known equation of state (cf. (3) or (4)).

An analytical solution of this system of differential equations for  $\vec{v}$  would not make much physical sense: these equations do not describe the physical reality completely. Also there is a lack of information about the boundary values. Thus only partial solutions, which take into account the observational evidence from the troposphere, are normally attempted. For instance, it is known that the vertical acceleration of the air in the troposphere is generally very small [CHAPMAN AND LINDZEN, 1970]. By neglecting the internal accelerations  $\vec{F}$  and limiting ourselves just to the horizontal component of  $\vec{v}$ , eqn. (17) reduces to

$$\dot{\vec{v}}_{\text{hor}} = -2(\vec{\omega} \times \vec{v})_{\text{hor}} - \frac{1}{\sigma_a} (\nabla p)_{\text{hor}}. \quad (9.20)$$

In a steady flow, which is prevalent in the atmosphere, the acceleration on the left-hand side disappears, and what is left is the equation describing the *geostrophic*

*wind.* Evidently the geostrophic wind corresponds to a situation in which the Coriolis and pressure gradient accelerations balance out, and where there are negligible temporal variations of temperature and pressure. The geostrophic wind is fairly representative of the real situation. According to PETTERSSEN [1969], geostrophic winds account for about 70% of the air motion over the seas. The rest is taken up mainly by accelerated motions, which will not be discussed in this book. The equations for the geostrophic wind can be written in a fairly simple manner. Selecting a local Cartesian, right-handed coordinate system with  $y$  pointing north and  $x$  pointing east, the reader can verify that the equations reduce to

$$v_N = \frac{1}{2\sigma_a \sin \phi} \cdot \frac{\partial p}{\partial x}, \quad v_E = \frac{-1}{2\sigma_a \omega \sin \phi} \cdot \frac{\partial p}{\partial y}, \quad (9.21)$$

where  $v_N, v_E$  are the horizontal wind velocities in south-to-north and west-to-east directions, and  $\phi$  is the latitude.

It is interesting to look at the global circulation pattern that reflects both the geostrophic and *thermal winds*. FIG. 12 shows the general features as observed on the Northern Hemisphere, according to ROSSBY [1941]. Note the effect of the Coriolis force: westward drift of the northbound flow and eastward drift of the southbound flow. (The interested reader is also advised to compare this figure with FIG. 3.) In reality, other regular and irregular circulations are superimposed on these global features.

The last phenomenon worth mentioning here is the *atmospheric tide*. It is a comprehensive term describing both the true tidal as well as other periodic deformations of air masses. It is mainly of thermal origin [CHAPMAN AND LINDZEN, 1970], and its predominant frequency is solar semidiurnal, i.e., the wave  $S_2$  (see §8.1). The magnitude of the pressure variation with this frequency, as estimated by HAURWITZ [1965], is shown in FIG. 13. Various other tidal frequencies are also present, but their smaller amplitudes are obscured by variations caused by other phenomena.

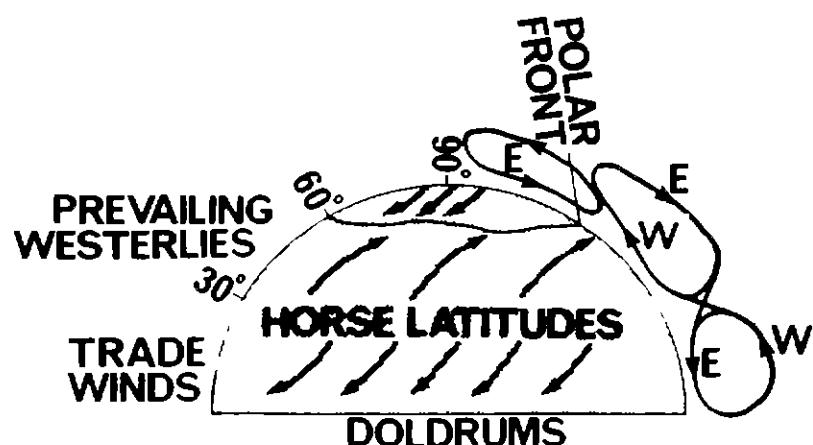


FIG. 9.12. Global circulation pattern for the Northern Hemisphere.

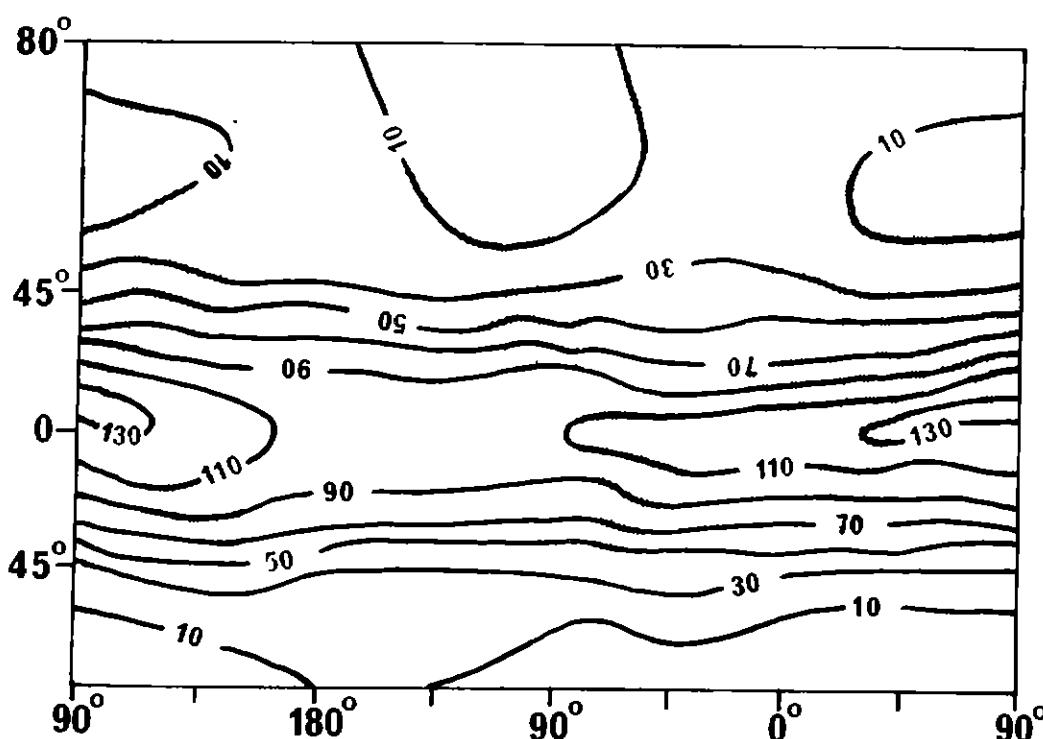


FIG. 9.13. Amplitude of solar semidiurnal atmospheric tide. Contours in microbars.

#### 9.4. Gravitational field of the atmosphere

Since the atmosphere has some mass, estimated to be about  $5.24 \times 10^{18}$  kg [COSPAR, 1965], it also gives rise to some gravitational attraction. The atmospheric mass is approximately  $0.89 \times 10^{-6}$  times smaller than that of the earth (cf. §6.1). As was seen earlier, gravitational acceleration and gravitational potential are both linear functions of the mass of the attracting body. Hence, the gravitational effects of the earth with the atmosphere, as observed by satellites outside the atmosphere, are 1.000 000 89 times larger than they would be for the earth without the atmosphere.

When investigating the earth's gravity field on and above the surface of the earth, it is often desirable (as will be shown in Part V) to be able to regard the space outside the earth as completely empty. It is thus advisable, for certain purposes, to correct the gravity observations made on the surface of the earth for the presence of the atmosphere. This correction must be considered when terrestrial and satellite results are compared. The *first-order atmospheric correction to gravity* would amount to adding 0.87 mGal to all the gravity observations made at sea level. The addition of such a correction would correspond to transferring (mathematically) the mass of the atmosphere to the centre of mass of the earth.

The *first-order atmospheric correction to the gravity potential*, translated into vertical displacement of the equipotential surfaces, is about 5 metres. This, however, has little meaning vis-à-vis the geoid because the geoid is defined by the mean sea level, as seen in §6.4. Thus this gravitational effect of the atmosphere can only uniformly change the value of the potential on the geoid.

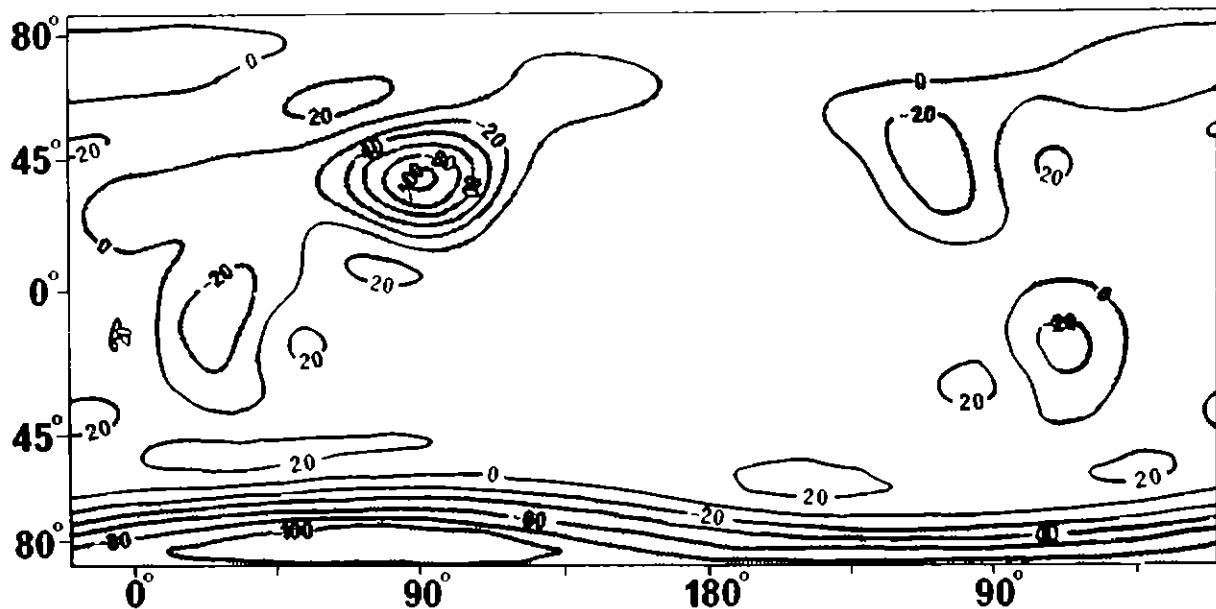


FIG. 9.14. Second-order atmospheric gravitational effect. Contours in microgals.

What can be done with gravity observations taken at a higher altitude? These should be corrected only for the effect of that part of the atmosphere above the point of observation; the effect of the layers below the point is reflected in the observed magnitude. The same correction also applies to airborne gravity observations. A table of such corrections, tabulated as a function of altitude, can be found in IAG [1971].

It can be shown that the gravitational potential of a spherical or ellipsoidal, laterally homogeneous shell is constant inside the shell [MACMILLAN, 1930]. Thus, if

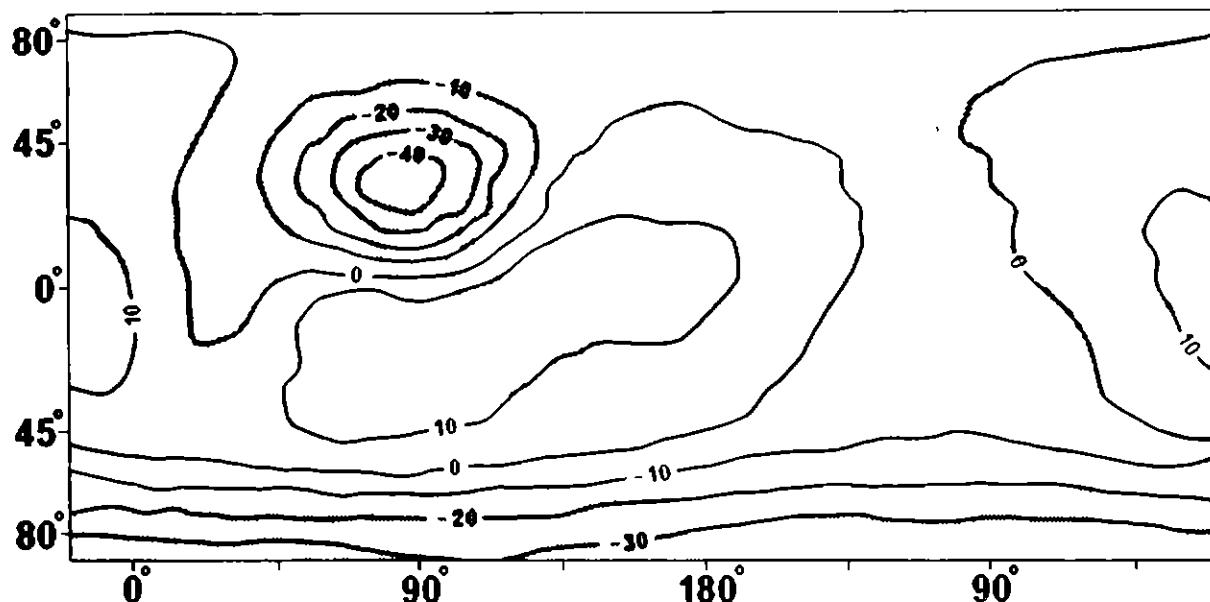


FIG. 9.15. Second-order atmospheric effect on the geoid determined from terrestrial gravity. Contours in centimetres.

the atmosphere could be regarded as consisting of homogeneous spherical or ellipsoidal strata, then all the points of the same altitude would experience the same atmospheric gravitational effect. This effect would be adequately described by the first-order correction discussed above. The real atmosphere is, however, neither laterally homogeneous nor regular in shape. In particular, its lower boundary is very irregular and, indeed, takes on the shape of the earth's topography. Because of this, the atmospheric gravitational effect is not the same even for points of equal altitude. This fact gives rise to the *second-order atmospheric correction to gravity*. The magnitude of this correction, according to ANDERSON ET AL. [1975], is shown in FIG. 14. In spite of the correction being generalized, the negative correlation with the topography is clearly visible.

The second-order effect would also change the potential irregularly, and thus it would result in a change of the shape of the geoid. One has to conclude, therefore, that the *second-order atmospheric correction to the gravity potential* may be applied to the shape of the geoid; it amounts to, at most, 10% of the first-order effect. Its magnitude, as computed by ANDERSON ET AL. [1975], is depicted in FIG. 15. Note again the negative correlation with the topography. The circumstances under which this correction, as well as the first-order correction, would be required depend on the intended usage of the geoid. More about this point will be said in §24.4.

## PART II

### REFERENCES

- AIRY, G.B. (1855). On the computations of the effect of the attraction of the mountain masses as disturbing the apparent astronomical latitude of stations in geodetic surveys. *Trans. Roy. Soc. London Ser. B* 145.
- ANDERLE, R.J. (1970). Polar motion determinations by U.S. Navy Doppler satellite observations. U.S. Naval Weapons Laboratory Technical Report 2432, Dahlgren, U.S.A.
- ANDERSON, E.G., C. RIZOS AND R.S. MATHER (1975). Atmospheric effects in physical geodesy. School of Surveying, Unisurv. No. G23, University of New South Wales, Kensington, Australia.
- ARNOLD, K. (1960). Numerische Beispiele zur strengen Theorie der Figur der Erde. Veröffentlichungen des Geodätischen Instituts Potsdam, Neue Serie Nr. 16, Germany.
- BÖHM, J. (1972). *Vyšší Geodesie I.* ČVUT, Prague, Czechoslovakia.
- BOWIE, W. AND H.G. AVERS (1914). Fourth general adjustment of the precise level net in the U.S. U.S. Coast and Geodetic Survey Special Publication 18, Washington, D.C., U.S.A.
- BROWN, R.D., S. VINCENT, W.E. STRANGE (1972). Undulation spectra for the marine geoid. Paper presented at the 53rd annual meeting of the American Geophysical Union, Washington, D.C., U.S.A.
- BUCHAR, E. (1958). Motion of the nodal line of the second Russian earth satellite and flattening of the earth. *Nature* 182, pp. 198–199.
- BULLEN, K.E. (1963). *Introduction to the Theory of Seismology*. 3rd ed., Cambridge University Press.
- BURNSIDE, C.D. (1971). *Electromagnetic Distance Measurement*. Crosby Lockwood.
- BURŠA, M. (1971). Fundamental geodetic parameters of the earth's figure and the structure of the earth's gravity field received from satellite data. Paper presented at the 15th General Assembly of the International Union of Geodesy and Geophysics, Moscow, U.S.S.R.
- CASSINIS, G. (1930). Sur l'adoption d'une formule internationale pour la pesanteur normale. *Bull. Géod.* 26, pp. 40–49.
- CESTONE, J.A., R.J. CYR, G. ROESLER AND E. ST. GEORGE JR. (1976). Latest highlights in acoustic underwater navigation. *Proc. International Navigational Congress*, U.S. Institute of Navigation, Boston Museum of Science, Boston, U.S.A., August, pp. 109–133.
- CHANDLER, S.C. (1891). On the variation of latitude. *Astronom. J.* 11, pp. 59–61, 65–70, 75–79, 83–86.
- CHAPMAN, S. AND R.S. LINDZEN (1970). *Atmospheric Tides*. Gordon and Breach.
- CLARKE, A.R. (1878). On the figure of the earth. *Philos. Mag. and J. Sci. London Ser. 5* 6, pp. 81–93.
- COMISION HIDROLOGICA DE LA CUENCA DEL VALLE DE MÉXICO (1961). Boletin de Mecanica de Suelos, No. 3, June 1956–June 1959. Secretaria de Recursos Hidraulicos, Oficina de Estudios Especiales, Mexico City, Mexico.
- CONDON, E.U. AND H. ODISHAW (EDS.) (1967). *Handbook of Physics*. 2nd ed., McGraw-Hill.
- COSPAR (1965). *COSPAR International Reference Atmosphere*. North-Holland.
- COULOMB, J. (1972). *Sea Floor Spreading and Continental Drift*. Reidel.
- CURRIE, R.G. (1974). Period and  $Q_w$  of the Chandler wobble. *Geophys. J. Roy. Astronom. Soc.* 38, pp. 179–185.

- DEPARTMENT OF ENERGY, MINES AND RESOURCES (1977a). Personal communication. Earth Physics Branch, Ottawa, Canada September.
- DEPARTMENT OF ENERGY MINES AND RESOURCES (1977b). Personal Communication. Earth Physics Branch, Ottawa, Canada September.
- DIEHL, W.S. (1948). Standard atmospheric tables and data. National Advisory Committee for Aeronautics, U.S.A.
- DOODSON, A.T. (1957). The analysis and prediction of tides in shallow water. *Internat. Hydrogr. Rev.* 33, pp. 85-126.
- DRUDE, P. (1959). *The Theory of Optics*. Dover.
- FAIRBRIDGE, R.W. AND O.A. KREBS (1962). Sea-level and the southern oscillation. *Geophys. J. Roy. Astronom. Soc.* 6 (4), pp. 532-545.
- FROST, N.H. AND J.E. LILLY (1966). Crustal movement in the Lake St. John area, Québec. *Canad. Surv.* 20 (4), pp. 292-299.
- GALE, L.A. (1970). Geodetic observations for the detection of vertical crustal movements. *Canad. J. Earth Sci.* 7, pp. 602-606.
- GAPOSHKIN, E.M. (1973). 1973 Smithsonian standard earth (III). Smithsonian Astrophysical Observatory Special Report 353, Cambridge, U.S.A.
- GARLAND, G.D. (1965). *The Earth's Shape and Gravity*. Pergamon.
- GASS, I.G., P.J. SMITH AND R.C.L. WILSON (EDS.) (1972). *Understanding the Earth*. 2nd ed., M.I.T. Press.
- GILBERT, G.K. (1890). *Lake Bonneville*. U.S. Geological Survey.
- GODIN, G. (1972). *The Analysis of Tides*. University of Toronto Press.
- GOLDRICH, P. AND A. TOOMRE (1969). Some remarks on polar wandering. *J. Geophys. Res.* 74 (10), pp. 2555-2567.
- GOUGH, D.I. AND W.I. GOUGH (1970). Stress and deflection in the lithosphere near Lake Kariba. *Geophys. J. Roy. Astronom. Soc.* 21, Part I, pp. 65-78, Part II, pp. 79-101.
- GRABER, M.A. (1976). Polar motion spectra based upon Doppler, IPMS, and BIH data. *Geophys. J. Roy. Astronom. Soc.* 46, pp. 75-85.
- GUINOT, B. (1977). Personal communication. Director of Bureau International de l'Heure, Paris, France. September.
- HAURWITZ, B. (1965). The diurnal surface pressure oscillation. *Arch. Met. Geophys. Biokl. A* 14, pp. 361-379.
- HAYFORD, J.F. (1909). The figure of the earth and isostasy from measurements in the United States. U.S. Coast and Geodetic Survey, Washington, D.C., U.S.A.
- HAYFORD, J.F. AND A.L. BALDWIN (1907). The earth movements in the California earthquake of 1906. U.S. Coast and Geodetic Survey Report for 1907, Append. 3, Washington, D.C., U.S.A.
- HEISKANEN, W.A. (1938). Investigations of the gravity formula. *Ann. Acad. Sci. Fenn. Ser. A* 51 (8), 22 pages.
- HEISKANEN, W.A. AND H. MORITZ (1967). *Physical Geodesy*. Freeman.
- HEISKANEN, W.A. AND F.A. VENING MEINESZ (1958). *The Earth and its Gravity Field*. McGraw-Hill.
- HELA, I. AND E. LISITZIN (1967). A world mean sea level and marine geodesy. *Proc. 1st Marine Geodesy Symposium*, Battelle Memorial Institute, Columbus, U.S.A., September, 1966. Government Printing Office, Washington, D.C., U.S.A., pp. 71-73.
- HENDERSHOTT, M.C. (1972). The effects of solid earth deformation on global ocean tides. *Geophys. J. Roy. Astronom. Soc.* 29, pp. 389-402.
- HENRIKSEN, S.W. (1960). The hydrostatic flattening of the earth. *Ann. of the IGY* 12 (1), pp. 197-198.
- HILL, M.N. (ED.) (1966). *The Sea*. Vol. I, Wiley Interscience.
- HIRVONEN, R.A. (1960). New theory of the gravimetric geodesy. Publications of the Isostatic Institute of the IAG, No. 32, Helsinki, Finland.
- HOLDAHL, S.R. (1969). Geodetic evaluation of land subsidence in the central San Joaquin Valley of California. In: *Reports on Geodetic Measurements of Crustal Movement, 1906-71*. U.S. Department of Commerce, Government Printing Office, Washington, D.C., U.S.A., 1973.
- HOLDAHL, S.R. AND N.L. MORRISON (1974). Regional investigations of vertical crustal movements in the U.S. using precise levellings and mareograph data. *Tectonophysics* 23, pp. 373-390.
- HOTINE, M. (1969). *Mathematical Geodesy*. ESSA Monograph 2. U.S. Department of Commerce, Government Printing Office, Washington, D.C., U.S.A.

- HYDROGRAPHER OF THE NAVY (1965). Admiralty manual of hydrographic surveying. Vol. I, Royal Navy, London, England.
- INTERNATIONAL ASSOCIATION OF GEODESY (1971). Geodetic reference system 1967. IAG Special Publication No. 3, Paris, France.
- INTERNATIONAL ASSOCIATION OF GEODESY (1980). The geodesist's handbook. Ed. I.I. Mueller, *Bull. Géod.* 54(3).
- INTERNATIONAL ASTRONOMICAL UNION (1977). *Proceedings of the Sixteenth General Assembly*. Ed. A. Muller, A. Jappel. IAU, Grenoble, 1976. Trans. of the IAU, Vol. XVIB, Reidel.
- IRVING, E. (1977). Drift of the major continental blocks since the Devonian. *Nature* 270, pp. 304–309.
- JEFFREYS, H. (1963). On the hydrostatic theory of the figure of the earth. *Geophys. J. Roy. Astronom. Soc.* 8, pp. 196–202.
- JEFFREYS, H. (1968). The variation of latitude. *Mon. Not. Roy. Astronom. Soc.* 141, pp. 255–268.
- JEFFREYS, H. (1970). *The Earth*. 5th ed., Cambridge University Press.
- KAULA, W. (1966a). Global harmonic and statistical analysis of gravity. In: *Extension of Gravity Anomalies to Unsurveyed Areas*, Ed. H. Orlin. American Geophysical Union Monograph 9, Washington, D.C., U.S.A., pp. 58–67.
- KAULA, W. (1966b). *Theory of Satellite Geodesy: Applications of Satellites to Geodesy*. Blaisdell.
- KAULA, W. AND J.A. O'KEEFE (1963). Stress differences and the reference ellipsoid. *Science* 142, p. 382.
- KOBOLD, F. AND E. HUNZIKER (1962). Communication sur la courbure de la verticale. *Bull. Géod.* 65, pp. 265–267.
- KOUBA, J. AND J.D. BOAL (1976). The Canadian Doppler satellite network. *Proc. International Geodetic Symposium on Satellite Doppler Positioning*, DMA and NOA of the NOAA, New Mexico State University, Las Cruces, U.S.A., October. Physical Science Laboratory of the New Mexico State University, pp. 187–206.
- KOVALEVSKY, J. (1967). *Introduction to Celestial Mechanics*. Vol. 7 in: "Astrophysics and Space Science Library." Translated by Express Translation Service. Springer/Reidel.
- KUKKAMÄKI, T.J. (1975). Report on the work of the Fennoscandian subcommission. *Proc. 4th International Symposium on the Problems of Recent Crustal Movements*, Ed. Yu. D. Boulanger. IUGG, Moscow, U.S.S.R., August. Valgus, Tallinn, pp. 25–29.
- LANDKOF, N.S. (1972). *Foundations of Modern Potential Theory*. Springer.
- LANGLEY, R.B., R.W. KING, I.I. SHAPIRO, R.D. ROSEN AND D.A. SALSTEIN (1981). Atmospheric angular momentum and the length of day: A common fluctuation with a period near 50 days. *Nature* 294 (5843), pp. 730–732.
- LEDERSTEGER, K. (1967). The equilibrium figure of the earth and the normal spheroid. *Analysed Proceedings of the International Symposium on the Figure of the Earth and Refraction*, Ed. K. Ledersteger. Austrian Geodetic Commission, Vienna, Austria, March, pp. 20–22.
- LEE, W.H.K. AND W. KAULA (1967). A spherical harmonic analysis of the earth's topography. *J. Geophys. Res.* 72, pp. 753–758.
- LE MÉHAUTÉ, B. (1976). *An Introduction to Hydrodynamics and Water Waves*. Springer.
- LEPICHON, X., J. FRANCHETEAU AND J. BONNIN (1973). *Plate Tectonics*. Elsevier.
- LIST, R.J. (ED.) (1958). Smithsonian meteorological tables. 6th ed., The Smithsonian Institution, Washington, D.C., U.S.A.
- MACMILLAN, D.H. (1966). *Tides*. CR Books.
- MACMILLAN, W.D. (1930). *The Theory of the Potential*. Dover reprint, 1958.
- MACMILLAN, W.D. (1936). *Dynamics of Rigid Bodies*. Dover reprint, 1960.
- MALONE, T.F. (ED.) (1951). Compendium of meteorology. American Meteorological Society, Boston, U.S.A.
- MANSINHA, L. AND D.E. SMYLIE (1967). Effects of earthquakes on the Chandler wobble and the secular polar shift. *J. Geophys. Res.* 72, pp. 4731–4743.
- MARKOWITZ, W. (1972) Rotational accelerations. *Proc. International Astronomical Union Symposium No. 48 on the Rotation of the Earth*, Eds. P. Melchior and S. Yumi. Morioka, Japan, May, 1971. Reidel, pp. 162–164.
- MARKOWITZ, W. AND B. GUINOT (EDS.) (1968). Continental drift, secular motion of the pole and rotation of the earth. *Proc. International Astronomical Union Symposium No. 32*, Stresa, Italy, March, 1967. Springer/Reidel.

- MCKEOWN, D.L. AND R.M. EATON (1974). An experiment to determine the repeatability of an acoustic range-range position system. *Proc. International Symposium on the Applications of Marine Geodesy*. Battelle Memorial Institute, Columbus, U.S.A., June, pp. 197–208.
- MCWHIRTER, N. AND R. MCWHIRTER (EDS.) (1975). *Guinness Book of World Records*. 13th ed., Bantam Books.
- Medallion World Atlas* (1973). Hammond Inc.
- MELCHIOR, P. (1966). *The Earth Tides*. Pergamon.
- MELCHIOR, P. (1972). *Physique et Dynamique Planétaires*. Vol. 3, Vander.
- MELCHIOR, P. (1973). *Physique et Dynamique Planétaires*. Vol. 4, Vander.
- MELCHIOR, P. (1978). *The Tides of the Planet Earth*. Pergamon.
- MENARD, H.W. (1973). Epirogeny and plate tectonics. *EOS, Trans. Am. Geophys. Union* 54 (12), pp. 1244–1255.
- MENZEL, D.H. (1955). *Fundamental Formulas of Physics*. Vol. 2, Dover reprint, 1960.
- MERRY, C.L. AND P. VANÍČEK (1974). A technique for determining the geoid from a combination of astrogeodetic and gravimetric deflection. *Canad. Surv.* 28 (5), pp. 549–554.
- MINZNER, R.A., K.S.W. CHAMPION AND H.L. POND (1959). The ARDC model atmosphere 1959. Air Force Surveys in Geophysics, No. 115, Washington, D.C., U.S.A.
- MISNER, C.W., K.S. THORNE AND J.A. WHEELER (1973). *Gravitation*. Freeman.
- MOLODENSKIY, M.S., V.F. EREMEEV AND M.I. YURKINA (1960). *Methods for Study of the External Gravitational Field and Figure of the Earth*. Translated from Russian by the Israel Program for Scientific Translations for the Office of Technical Services, Department of Commerce, Washington, D.C., U.S.A., 1962.
- MORITZ, H. (1973). Ellipsoidal mass distribution. Department of Geodetic Science Report 206, The Ohio State University, Columbus, U.S.A.
- MUELLER, I.I. (1969). *Spherical and Practical Astronomy as Applied to Geodesy*. Ungar.
- MUELLER, I.I. (Ed.) (1975). *Proceedings of the Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conferences, 1972–1974*. Department of Geodetic Science Report 231, The Ohio State University, Columbus, U.S.A.
- MUNK, W.M. AND G.F.K. MACDONALD (1960). *The Rotation of the Earth*. Cambridge University Press.
- NAGY, D. (1973). Free air anomaly map of Canada from piece-wise surface fittings over half-degree blocks. *Canad. Surv.* 27 (4), pp. 293–300.
- NAKAGAWA, J., P. MELCHIOR AND H. TAKEUCHI (1968). Free oscillations of the earth observed by a gravimeter at Brussels. Observatoire Royale Belge Communication #1, Brussels, Belgium.
- NASSAU, J.J. (1948). *Practical Astronomy*. 2nd ed., McGraw-Hill.
- NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (1973). NASA directory of observation station locations. Vols. 1 and 2, 3rd ed., Goddard Space Flight Center, Greenbelt, U.S.A.
- NEWCOMB, S. (1892). Remarks on Mr. Chandler's law of variation of terrestrial latitudes. *Astronom. J.* 12, pp. 49–50.
- NEWCOMB, S. (1906). *A Compendium of Spherical Astronomy*. Dover reprint, 1960.
- NEWTON, I. (1687). *Philosophiae Naturalis Principia Mathematica*. 3rd ed., Cambridge University Press reprint, 1972.
- OFFICER, C.B. (1974). *Introduction to Theoretical Geophysics*. Springer.
- ORLOV, A.YA. (1961). §15, Izbranye Trudy. Sluzhba Shiroty, Kiev, U.S.S.R. (English translation: §15, Collected works. Latitude Service of the U.S.S.R. Kiev).
- PEDERSEN, G.P.M. AND M.G. ROCHESTER (1972). Spectral analysis of the Chandler wobble. *Proc. International Astronomical Union Symposium No. 48 on the Rotation of the Earth*, Eds. P. Melchior and S. Yumi. Morioka, Japan, May, 1971. Reidel, pp. 33–38.
- PELIKAN, M. (1967). The calculation of refraction angles by means of the refractive index and the radii of curvature of the refractive curve. *Analysed Proceedings of the International Symposium on the Figure of the Earth and Refraction*, Ed. K. Ledersteger. Austrian Geodetic Commission, Vienna, Austria, March, pp. 211–219.
- PELTIER, W.R., W.E. FARRELL AND J.A. CLARK (1978). Glacial isostasy and relative sea level: A global finite element model. *Tectonophysics* 50, pp. 81–110.
- PERMANENT SERVICE FOR MEAN SEA LEVEL (1976). Monthly and annual mean heights of sea level. Vol. 1, Institute of Oceanographic Sciences, Birkenhead, England.

- PETTERSEN, S. (1969). *Introduction to Meteorology*. 3rd ed., McGraw-Hill.
- PICK, M., J. PÍCHA AND V. VYSKOČIL (1973). *Theory of the Earth's Gravity Field*. Elsevier.
- POLAND, J.F. AND G.H. DAVIS (1969). Land subsidence due to withdrawal of fluids. Reprinted from *Reviews in Engineering Geology*: II, pp. 187–269, by The Geological Society of America Inc., Boulder, U.S.A.
- PRATT, J.H. (1855). On the attraction of the Himalaya Mountains and of the elevated regions beyond upon the plumb-line in India. *Trans. Roy. Soc. London Ser. B* 145, pp. 53–100.
- PREY, A. (1922). Darstellung der Höhen- und Tiefenverhältnisse der Erde durch eine Entwicklung nach Kugelfunctionen bis zur 16 Ordnung. *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen* II, Part 1.
- QURAISHEE, G.S. AND P. VANÍČEK (1970). A search for low frequencies in residual tide and mean sea level observations by means of the least-squares spectral analysis. *Report on the Symposium on Coastal Geodesy*, Ed. R. Sigl. IUGG, IAG, Munich, West Germany, July. Institut für Angewandte Geodäsie, pp. 485–493.
- RAPP, R.H. (1974). The geoid: Definition and determination. *EOS, Trans. Am. Geophys. Union* 55 (3), pp. 118–126.
- RIKITAKE, T. (1976). *Earthquake Prediction*. Elsevier.
- ROCHESTER, M.G. (1973). The earth's rotation. *EOS, Trans. Am. Geophys. Union* 54 (8), pp. 769–781.
- ROSSBY, C.G. (1941). The scientific basis of modern meteorology. In: *Climate and May Yearbook of Agriculture*, U.S. Department of Agriculture, Washington, D.C., U.S.A.
- RUNCORN, S.K. (ED.) (1967). *International Dictionary of Geophysics*. Vol. I, Pergamon.
- SAVAGE, J.C. AND R.O. BURFORD (1973). Geodetic determination of relative plate motion in central California. *J. Geophys. Res.* 78 (5), pp. 832–845.
- SCHMID, H.H. (1974). Worldwide geocentric satellite triangulation. *J. Geophys. Res.* 79 (35), pp. 5349–5376.
- SCHWARZ, K.P. (1975). Zur Erdmessung des Eratosthenes. *Allgem. Vermessungs-Nachr.* 82 (1), pp. 1–12.
- SEPELIN, T.O. (1974). The Department of Defense world geodetic system 1972. *Canad. Surv.* 28 (5), pp. 496–506.
- SIMMONS, L.G. (1950). How accurate is first-order triangulation? *J. U.S. Coast and Geod. Surv.* 3, pp. 53–56.
- SMART, W.M. (1956). *Text-book on Spherical Astronomy*. Cambridge University Press.
- STOMMEL, H. (1963). Varieties of oceanographic experience. *Science* 139, pp. 572–576.
- SYMON, K.R. (1971). *Mechanics*. 3rd ed., Addison-Wesley.
- TERRIEN, J. (1974). International agreement on the value of the velocity of light. *Metrologia* 10 (9).
- TSUBOI, C. (1933). Investigation on the deformation of the earth's crust found by precise geodetic means. *Jap. J. Astronom. Geophys.* 10, pp. 93–248.
- U.S. ARMY TOPOGRAPHIC COMMAND (1971a). Fundamental geodetic networks (vertical control). Department of Defense, Washington, D.C., U.S.A.
- U.S. ARMY TOPOGRAPHIC COMMAND (1971b). Fundamental geodetic networks (horizontal control). Department of Defense, Washington, D.C., U.S.A.
- VANÍČEK, P. (1969). New analysis of the earth pole wobble. *Stud. Geoph. et Geod.* 13, pp. 225–230.
- VANÍČEK, P. (1971). An attempt to determine long-periodic variations in the drift of horizontal pendulums. *Stud. Geoph. et Geod.* 15, pp. 416–420.
- VANÍČEK, P. AND A.C. HAMILTON (1972). Further analysis of vertical crustal movement observations in the Lac St. Jean area, Québec. *Canad. J. Earth Sci.* 9 (9), pp. 1139–1147.
- VANÍČEK, P. AND C.L. MERRY (1973). Determination of the geoid from deflections of the vertical using a least-squares surface fitting technique. *Bull. Géod.* 109, pp. 261–279.
- VENING MEINESZ, F.A. (1931). Une nouvelle méthode pour la réduction isostatique régionale de l'intensité de la pesanteur. *Bull. Géod.* 29, pp. 33–45.
- VINCENT, S., W.E. STRANGE AND J.G. MARSH (1972). A detailed gravimetric geoid from North America to Eurasia. Goddard Space Flight Center Report X-553-72-94, Greenbelt, U.S.A.
- WALCOTT, R.I. (1972). Late quaternary vertical movements in eastern North America: Quantitative evidence of glacio-isostatic rebound. *Rev. Geophys. and Space Phys.* 10 (4), pp. 849–884.
- WALCOTT, R.I. (1975). Recent and late quaternary changes in water level. *EOS, Trans. Am. Geophys. Union* 56 (2), pp. 62–72.

- WEGENER, A. (1929). *The Origin of Continents and Oceans*. Translation of 1962 printing of 4th rev. German ed., Friedr Vieweg and Sohn, Dover reprint, 1966.
- WEIFFENBACH, G.C. (1967). Tropospheric and ionospheric propagation effects on satellite radio-Doppler geodesy. *Proc. Symposium on Electromagnetic Distance Measurements*. Oxford, England, September, 1965. University of Toronto Press, pp. 339-352.
- WELLS, D.E. (1979). Personal communication. Department of Surveying Engineering, University of New Brunswick, Fredericton, Canada.
- WELLS, F.J. AND M.A. CHINNERY (1973). On the separation of the spectral components of polar motion. *Geophys. J. Roy. Astronom. Soc.* 34, pp. 179-192.
- WHITTEN, C.A. (1970). Crustal movement from geodetic measurements. *Proc. NATO Advanced Study Institute Conference on Earthquake Displacement Fields and the Rotation of the Earth*, Eds. L. Mansinha, D.E. Smylie and A.E. Beck. University of Western Ontario, London, Canada, June, 1969. Springer/Reidel, pp. 255-268.
- WILL, L.S. (ED.) (1971). *Studies of Space Experiments to Measure Gravitational Constant Variations and Eötvös Ratio*. M.I.T. Press.
- WILSON, G. AND H. GRACE (1942). The settlement of London due to underdrainage of the London clay. *J. Inst. Civ. Eng.* 19 (2), pp. 100-127.
- World Almanac and Book of Facts 1984, The (1983). Newspaper Enterprise Association, Inc.
- YERKES, R.F. AND R.O. CASTLE (1971). Surface deformation associated with oil and gas field operations in the U.S. *Proc. Conference on Land Subsidence*, Ed. L.J. Tison. Tokyo, Japan, 1969. Association Internationale d'Hydrologie Scientifique n. 89, pp. 55-66.
- YUMI, S. (1970). Polar motion in recent years. *Proc. NATO Advanced Study Institute Conference on Earthquake Displacement Fields and the Rotation of the Earth*, Eds. L. Mansinha, D.E. Smylie and A.E. Beck. University of Western Ontario, London, Canada, June, 1969. Springer/Reidel, pp. 45-53.
- YUMI, S. (1977). Personal communication. International Polar Motion Service, Mizusawa, Japan. September.

## **PART III**

# **METHODOLOGY**

## CHAPTER 10

### ELEMENTS OF GEODETIC METHODOLOGY

In this opening chapter of Part III, an overview is given of the methodology used in geodesy. The first section describes the general procedure normally adhered to when performing a geodetic task. This is followed by a section dealing with the general principles involved in the formulation of mathematical models. The last two sections are concerned with the characteristics of observable quantities, i.e., those quantities through which the unknown parameters are determined. First, the properties of an isolated observable are given, and then the treatment of vectors of observables is outlined. These last two sections serve as the basis for the eventual formalization of the treatment of least-squares problems. It was not deemed appropriate to the scope of this book to treat either instrumentation or measuring techniques here. For these the reader is referred to other sources cited in the text.

#### 10.1. General procedure

*Geodetic methodology* is a set of procedures adopted for the evaluation of quantities that contribute directly or indirectly to the description of the geometry of the earth and its gravity field. Every experiment or project should be designed around specifications placed on the quantities that are being investigated. The design thus comprises the determination of the kind and amount of data that need be collected, as well as their accuracies. These data are then procured, screened, and analysed to see whether they actually fulfil the prescribed accuracy specifications. Once scrutinized, these data are processed, and solutions are obtained for the quantities of interest. Finally, the results are evaluated and presented. Although the methodology used in geodesy is similar to other experimental sciences certain circumstances in geodesy place constraints upon both the kind and the number of assumptions that can be made and thereby dictate the specific procedures that must be adopted.

The most significant factor affecting the practice of geodesy, as with other sciences, is economics. Often, geodetic operations involve expensive instrumentation and extensive field operations that result in large expenditures. For example, the establishment of a single geodetic network (see §7.1) of continental extent (whether it be a gravity, horizontal, or height network) can cost tens of millions of dollars. Thus the remeasurement of networks cannot be undertaken as readily as can some laboratory experiments. Consequently, to maximize the return on the investment,

geodetic procedures must include optimization of design and planning, careful collection of data, meticulous assessment of the collected data, and rigorous evaluation of the results. Another factor peculiar to geodesy is that usually more data is collected than is needed for a unique determination of the desired quantities. This is done on purpose to have a means of assessing the accuracy and reliability of the results.

In geodesy, the mathematical models relating the collected data to certain unknown parameters are fairly well known, because they are based on geometrical and simple physical laws. This contrasts, for instance, the social sciences where, because the mathematical model underlying the data is often unknown, special techniques need be utilized to help in the creation of a model. Further, models used in geodesy are often non-linear and thus require use of somewhat involved mathematical techniques in their solutions.

The above should serve as the basis for a better appreciation of the following stages that constitute the geodetic methodology:

(a) At the outset, the quantities of interest, called unknown parameters, are identified, and the desired accuracies of these parameters are usually prescribed. This must be done on the basis of an intimate knowledge of the project and there is no preset way of doing it.

(b) Because the unknown parameters generally cannot be measured directly, it is necessary to formulate functions that relate them to other quantities that can be measured—the observables. Consequently, the second stage concerns the formulation of these functions, called a *mathematical model*, that is the basis for the determination of the unknown parameters. Formulation of mathematical models is really what this book should prepare the reader to do.

(c) Before making the required measurements, their accuracy must be specified. Clearly, these accuracies are dictated by the desired accuracy of the unknown parameters and by the formulated mathematical model. This optimum accuracy design has become known as *preanalysis*, and the selection of the measuring procedure is based on it. The problem of preanalysis is treated in §14.1.

(d) Individual measurements are then made and observations are assessed (screened) to find out whether or not they meet the prescribed accuracy specifications. If not, remeasurement is required. While the task of making measurements is not discussed in this book, observation assessment is covered in §13.3.

(e) The preprocessed observations are then introduced into the mathematical model, and the unknown parameters and their accuracies are solved. Although the numerical techniques for actually solving the involved systems of equations are not a topic of this book, the ways of transferring mathematical models to such solvable systems are shown in Chapters 11 and 12.

(f) The next stage is the assessment of the mathematical model for completeness together with a further assessment of the observations for correctness. This task is treated in detail in §13.4.

(g) The final stage involves an assessment of the computed unknown parameters and an examination of their compatibility with other independent determinations of the same parameters, should these exist. Procedures used at this stage are described in §13.5, and some special techniques are scattered throughout the book.

## 10.2. Formulation of the mathematical model

The formulation of the functional relation between the unknown parameters and the observed quantities plays a key role in geodetic methodology. The model is the central element in both designing the experiment and processing the observed data. The mathematical model is simply a mathematical relation between particular quantities that is based on certain laws. In symbolic form, the mathematical model can be written as

$$\mathbf{f}(\mathbf{q}) = \mathbf{0}, \quad (10.1)$$

where  $\mathbf{f}$  denotes the vector of individual functions  $f_i$ ,  $i = 1, \dots, m$ , that link together  $N$  quantities  $q_i$ , denoted by the vector  $\mathbf{q}$ . Because of the involvement of the laws of nature or geometry, some of the constituents of  $\mathbf{q}$  can be considered as completely known or, in statistical terminology, as errorless. These quantities are also known as *constants* and are denoted by the vector  $\mathbf{c}$ . Examples of such constants are Newton's gravitational constant (eqn. (6.1)), the sum of angles in a plane triangle, or the velocity of light in a vacuum (see §9.2). Generally, such fundamental constants are assumed known, as it is not the goal of geodesy to improve upon their values.

In contrast to constants, there are quantities for which there is little or no information. These are the *unknown parameters*,  $x_i$ ,  $i = 1, \dots, u$ , denoted by the vector  $\mathbf{x}$ . Throughout this book, the parameters are assumed to be mutually independent quantities; i.e., direct computation of any one of these parameters from the others is impossible except when it is clear from the text that constraints are imposed (see later in this section). Some examples of parameters are heights or other coordinates, the deflections of the vertical, geoidal heights, and time variations of coordinates.

Falling somewhere between the constants and unknown parameters are the quantities called *observables*. As stated earlier, an observable is a physical or geometrical quantity that can be observed, i.e., a quantity to which a number can be assigned with a certain accuracy. An *observation*  $l_i$ ,  $i = 1, \dots, n$ , is the number assigned to the observable. The process of assigning this number is termed the *measurement* and is accomplished by means of an *instrument* or *sensor*.

The mathematical model (1) can now be rewritten, in terms of constants  $\mathbf{c}$ , parameters  $\mathbf{x}$ , and observations  $\mathbf{l}$ , as

$$\mathbf{f}(\mathbf{q}) = \mathbf{f}(\mathbf{c}, \mathbf{x}, \mathbf{l}) = \mathbf{0}, \quad (10.2)$$

where, clearly, the vector  $\mathbf{q}$  has been partitioned into three parts. From this point on, the explicit reference to the existence of the constants  $\mathbf{c}$  in the model will be omitted, and it will be understood that  $\mathbf{c}$  is part of the function (model) itself. Further, it will be understood that *unknown parameters*, in general, cannot be measured directly. They are usually indirectly determined, through the mathematical model, by the observations. For this reason,  $\mathbf{x}$  is also often called the *solution*, meaning the solution to whatever problem is dealt with. Observations that are not functionally related to any unknown parameters are useless.

Corresponding to the three components of the mathematical model (2) are the three mathematical spaces: parameter, observation, and model space (FIG. 1—the

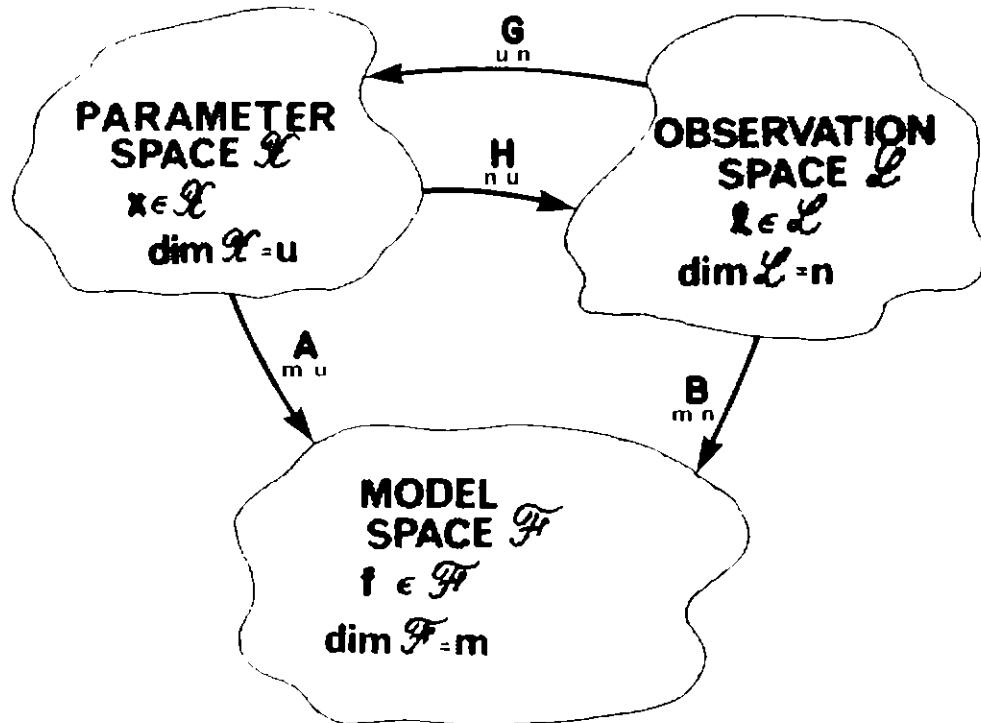


FIG. 10.1. Linear relations between parameter, observation, and model spaces.

matrices  $A$ ,  $B$ ,  $G$ ,  $H$ , and their roles, will be defined later in this chapter). *Parameter space*, or *solution space*, is defined as the set of all possible  $x$ 's and is denoted by  $\mathcal{X}$ ;  $\dim \mathcal{X} = u$ . *Observation space* is defined as the set of all possible  $l$ 's and is denoted by  $\mathcal{L}$ ;  $\dim \mathcal{L} = n$ . *Model space* is defined as the set of all possible  $f$ 's and is denoted by  $\mathcal{F}$ ;  $\dim \mathcal{F} = m$ .

Models may be of the direct, indirect, or implicit variety; may be linear or non-linear; and may exist alone or in combinations. In the case of a grouping of several models, some models play a primary role, while others play only a secondary role.

(a) The *model explicit in  $x$*  is written as

$$x = g(l), \quad (10.3)$$

where  $g$  is an explicit function. Because  $g$  transforms  $\mathcal{L}$  into  $\mathcal{X}$ , both sides of (3) belong to  $\mathcal{X}$  and the model is said to be formulated in parameter space  $\mathcal{X}$ . The linear version of the explicit form is the *linear model explicit in  $x$*

$$x = Gl + w, \quad (10.4)$$

where  $G$ , the matrix transforming  $\mathcal{L}$  into  $\mathcal{X}$  (cf. FIG. 1), is called the *design matrix*. Its elements, and those of the *constant vector  $w$* , are known, because both  $G$  and  $w$  must reflect the known physical or geometrical design of the experiment;  $\dim G =$

$(u, n)$ ,  $\dim w = u = m$ . The following linear explicit model is an example:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 \\ -3 & 4 & 5 \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \\ l_3 \end{bmatrix} + \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

The trivial case of the explicit model occurs when it is possible to make a direct measurement of the unknown parameters. In this situation, (4) takes the form

$$x = l, \quad (10.5)$$

so that  $G = I$ , where  $I$  is the unit matrix (see §3.1);  $w = 0$ ; and  $u = n = m$ .

Under certain circumstances, the unknown parameters may be totally absent in the explicit model. In this case, the model becomes

$$g(l) = 0, \quad (10.6)$$

and is known as the *condition model* reflecting the physical or geometrical conditions relating only the observables among themselves. Its linear version is the *linear condition model*

$$Gl + w = 0, \quad (10.7)$$

where  $\dim G = (m, n)$ ,  $\dim l = n$ , and  $\dim w = m$ . For example, in dealing with the three angles,  $\alpha, \beta, \gamma$ , of a plane triangle, the design matrix is

$$G = [1, 1, 1],$$

the vector of observables is

$$l = [\alpha, \beta, \gamma]^T,$$

and the constant vector of one element in this case equals

$$w = [-\pi].$$

(b) Often it is easier to express the observations as functions of the parameters rather than vice versa. These (explicit) functions represent the *model explicit in l*; namely,

$$l = h(x). \quad (10.8)$$

Because  $h$  transforms  $\mathcal{X}$  into  $\mathcal{L}$ , this type of model is said to be formulated in the observation space. As an illustration of this model, consider the following:

$$\begin{bmatrix} l_1 \\ l_2 \\ l_3 \end{bmatrix} = \begin{bmatrix} h_1(x_1, x_2) \\ h_2(x_1, x_2) \\ h_3(x_1, x_2) \end{bmatrix} = \begin{bmatrix} 3x_1^2 + \cos x_2 \\ x_1 - \sin x_2 \\ x_1^2 + x_2 \end{bmatrix}.$$

In this example, it is obvious that for each observation  $l_i$ , there is one equation  $h_i$ , involving the two unknown parameters  $x_1$  and  $x_2$ . As there are more equations than

unknowns, the model is said to be overdetermined. If there are fewer equations than unknowns, then the model is underdetermined, and if the number of equations and the number of unknowns are equal, then the model is uniquely determined. The treatment of these models is the subject of Chapter 11.

If the observations can be written as linear functions of the parameters, then we speak of the *linear model explicit in  $\mathbf{l}$* : namely,

$$\mathbf{l} = \mathbf{H}\mathbf{x} + \mathbf{w}, \quad (10.9)$$

where  $\mathbf{H}$  is, again, a design matrix transforming  $\mathcal{X}$  into  $\mathcal{L}$  (cf. FIG. 1) and  $\mathbf{w}$  is a vector containing known elements;  $\dim \mathbf{H} = (n, u)$ ;  $\dim \mathbf{w} = n = m$ .

(c) In some instances, the observables and unknown parameters have an interwoven relation. Such a form is known as the *implicit model*; in equation form it is written as

$$\mathbf{f}(\mathbf{x}, \mathbf{l}) = \mathbf{0}. \quad (10.10)$$

The above formulation is in the model space with the elements being the  $m$  functions;  $\dim \mathbf{f} = m$ . Following is a non-linear example, where  $m = n = u = 2$ :

$$\begin{bmatrix} f_1(l_1, l_2, x_1, x_2) \\ f_2(l_1, l_2, x_1, x_2) \end{bmatrix} = \begin{bmatrix} l_1^2 \sin x_1 + (l_2 x_2)^{1/2} \\ l_1 \exp(x_2) + l_2 x_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The *linear implicit model* is written as:

$$\mathbf{Ax} + \mathbf{Bl} + \mathbf{w} = \mathbf{0}, \quad (10.11)$$

where  $\mathbf{A}$  is called the *first design matrix* and  $\dim \mathbf{A} = (m, u)$ ;  $\mathbf{B}$  is called the *second design matrix* and  $\dim \mathbf{B} = (m, n)$ ; while  $\mathbf{w}$  is, again, the known constant vector and  $\dim \mathbf{w} = m$ . The matrices  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{w}$  are known from the design of the experiment; note that  $\mathbf{A}$  transforms  $\mathcal{X}$  into  $\mathcal{F}$ , and  $\mathbf{B}$  transforms  $\mathcal{L}$  into  $\mathcal{F}$  (cf. FIG. 1). The dimensions of the matrices  $\mathbf{A}$  and  $\mathbf{B}$ , and their ranks, dictate whether the model is underdetermined, overdetermined, or uniquely determined (see Chapter 11). Clearly, the implicit model is the most general, and the two explicit forms, (a) and (b), are only special cases of it.

Let us now turn our attention to models through which it is possible to introduce auxiliary information, i.e., information not contained in the primary models  $\mathbf{g}$ ,  $\mathbf{h}$ , and  $\mathbf{f}$ . This information is in the form of additional functional relations among the parameters or observables themselves. In the literature, these are known as *constraint functions* or models. A constraint model for parameters alone is written as

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}, \quad (10.12)$$

while one in terms of the observables is given by (6). Constraint models containing both the parameters and observables are special cases of the implicit model given by (10). Constraint models may, again, be linear or non-linear. With the exception of the condition model, constraint models are not treated alone, as are the models discussed earlier, but coexist with other models as exemplified by the following

*combination of models:*

$$\begin{aligned} f(\mathbf{x}, t) &= \mathbf{0}, \\ h(\mathbf{x}) &= \mathbf{0}, \end{aligned} \tag{10.13}$$

where  $f$  and  $h$  are known respectively as the *primary* and *secondary models*. The solution of such a combination of models—non-linear or linear—is the particular subject of §14.5.

### 10.3. Observables and their properties

There are a multitude of physical and geometrical quantities in geodesy that can be classified as observables. *Horizontal angles* and *directions* are two of the most common observables and are usually measured with a theodolite [COOPER, 1974]. Other common observables are distances between terrestrial points, whether they be *horizontal distances* obtained with a subtense bar [SMITH, 1970; HODGES AND GREENWOOD, 1971] or *spatial distances* measured by a tape, wire, or electronic distance measuring (EDM) equipment [BURNSIDE, 1971]. Levelled *height differences* and *vertical angles* are other examples of conventional, terrestrial observables [BOMFORD, 1971]. Measurements to objects in space are rapidly becoming very important in geodesy. Distances and directions are now routinely measured from the earth to the moon and to earth's artificial satellites. A *direction to a satellite* can be obtained, for instance, by photographing the satellite against the star background using special satellite cameras [VEIS, 1963; MUELLER, 1964]. A *distance to a satellite* is obtained by accurately timing the propagation of electromagnetic waves emitted by a source such as a laser [LEHR ET AL., 1974]. It is also possible to measure the *distance difference* from a ground station to two contiguous satellite positions by utilizing the change in the satellite-emitted frequency caused by the Doppler effect [GUIER AND WEIFFENBACH, 1960]. *Vertical and horizontal angles to stars* are usually obtained by means of special theodolites [MUELLER, 1969; ROBBINS, 1976].

Observables pertaining to the earth's gravity field are of vital importance to geodesy. Two such observables, *gravity* and *gravity differences*, are measured by instruments ranging from gravimeters [COOK, 1973] and pendulums to apparatuses utilizing a body falling in a vacuum chamber [FALLER, 1965]. *Gradients of gravity* are obtained using an instrument known as a torsion balance [MUELLER, 1963] or by gradiometers [FORWARD, 1974].

Observables that vary with time help to determine the change in the geometry of the earth. An example of such an observable is *sea level variations*. Recently, submersible tide gauges have enhanced the collection of this kind of data formerly confined to shore areas by the use of land based gauges [LENNON, 1970; 1974]. *Distance variation* has also become an important observable; it can be measured with strain gauges [VALI ET AL., 1965], or other instruments. *Variations of tilt*—another observable—most frequently employs a horizontal pendulum as the sensing apparatus [MELCHIOR, 1978], but other designs exist. Measurement of *time* itself, either as a part of a measuring apparatus or in connection with determining the epoch of observations to extraterrestrial objects is, of course, essential.

To completely understand the concept of an observable, the notions of time and space, and the measurement of one or several observables in both time and space, must be introduced. This means that a single observable, e.g., a distance between a pair of terrestrial points, can be measured over and over to yield a *series of observations in time*. On the other hand, distances between different pairs of points in a network can be measured to give a *series of observations in space*. In addition, this discussion must not be restricted to only one kind of observable, but must make allowance for the simultaneous treatment of several kinds of observables, e.g., distances mixed with directions.

Let us now express the above mathematically by first considering only one kind of observable,  $l$ , measured in time  $\tau$  to give observations

$$l(\tau_i), \quad i = 1, \dots, N. \quad (10.14)$$

The above expression represents, e.g., a distance measured between a pair of points at  $N$  different instants  $\tau_i \in \mathcal{T} \equiv \{\tau_1, \tau_2, \dots, \tau_N\}$ . (The parameter  $\tau$  may also denote a spatial rather than time coordinate of the observation  $l(\tau_i)$ . If such a distinction is important, it will be clear from the context which interpretation is intended.) The values  $\tau_i$  of the time (or space) parameter  $\tau$ , for which the observations are made, are called *sample points*. In the example above, the sample points  $\tau_i$  correspond to the instants at which the distance was measured or, more rarely, to the location of a distance in the network.

The case of several observables, possibly of different kinds and, typically, separated in space, will be denoted as

$$\mathbf{l} = [l_1, l_2, \dots, l_n]^T, \quad (10.15)$$

where each component  $l_i$  of this vector is a result of consolidating the  $N_i$  elements of the series given by (14) into one *representative value of the observable*. Before consolidation, there must, therefore, be a *table of observations*; namely,

$$\mathbf{L} = \begin{pmatrix} l_1(\tau_1^1) & l_1(\tau_2^1) & \dots & l_1(\tau_{N_1}^1) \\ l_2(\tau_1^2) & l_2(\tau_2^2) & \dots & l_2(\tau_{N_2}^2) \\ \vdots & & & \\ l_n(\tau_1^n) & l_n(\tau_2^n) & \dots & l_n(\tau_{N_n}^n) \end{pmatrix} \quad (10.16)$$

In practice, each row  $(l_i(\tau_j^i), \quad i = 1, \dots, N_i)$  of this table is replaced by one representative value, and the resultant column vector  $\mathbf{l}$  is then considered. Two points need be made: First,  $l_1(\tau_j^1)$  was, in all probability, measured at a time different from  $l_2(\tau_j^2)$ , i.e., the values of  $\tau_j$  in each column are generally not identical:  $\tau_j^i \neq \tau_j^k$  when  $j \neq k$ . Second, the number of elements in each row is generally different.

How is a representative value of  $l$  for the particular row of  $\mathbf{L}$  obtained? This is achieved in three steps by

- (a) modelling  $l(\tau)$ ;
- (b) analysing the fit of the model to  $l(\tau)$ ; and
- (c) computing a value of  $l$ , which may or may not correspond to a specific value of  $\tau$ .

The value determined in this way is then used in the vector of observables  $\mathbf{l}$ . The simplest situation would allow the representative value to be computed outright from a simple auxiliary model, such as the arithmetic mean of all the  $l(\tau)$ 's.

As the *auxiliary model* is generally unknown, one must be hypothesized. Any such hypothesis provides a *deterministic model* which, in all likelihood, will not completely describe the variation of  $\mathbf{l}$  with  $\tau$ . A *stochastical model* is then used to account for this lack of fit. It is thus expedient to think of  $l(\tau)$  as generally containing both the deterministic and the stochastical (random) components (see §3.4) which are referred to as the *trend*  $t$  and the *residual*  $r$  (the negative value of which is often called deviation); namely,

$$l(\tau) = t(\tau) - r(\tau). \quad (10.17)$$

The word ‘residual’ is to be understood as ‘residual of  $t$  after  $l$  has been subtracted’.

The trend, in turn, can be decomposed as follows:

$$t(\tau) = \hat{t}(\tau) + p(\tau), \quad (10.18)$$

where  $\hat{t}(\tau)$  denotes the *expected value of the observable* (the exact meaning of the word ‘expected’ will be made clear in the proper context (§13.1)), and  $p(\tau)$  is the *systematic component*, i.e., that variation of the observable that can be expressed through a formula containing certain parameters. These parameters characterize the systematic trends affecting the observations that have not been successfully removed through the measurement process. The systematic component is, thus, written as the linear combination

$$p(\tau) = \tilde{\Phi}(\tau)\lambda, \quad (10.19)$$

where  $\tilde{\Phi}(\tau)$  is one column of the Vandermonde matrix (see §3.1). The functions used in this matrix are referred to as base functions, and more will be said about them in §14.2. In the literature, the quantities  $\lambda$  are sometimes called *nuisance parameters*. For example, when processing EDM measurements, it may be advantageous to introduce nuisance parameters to model and remove the unaccounted for refraction.

If the auxiliary model is suspected of being incomplete, then systematic components (errors) can be expected to contaminate the preprocessed data. With this suspicion and a knowledge of the problem, an attempt to account for the effect of these systematic errors further down the line, i.e., within the context of the primary mathematical model, can be made by seeking, in addition to the unknown parameter vector  $\mathbf{x}$ , some nuisance parameters  $\lambda$ .

The stochastical residual may be decomposed as follows:

$$r(\tau) = v(\tau) + s(\tau), \quad (10.20)$$

where  $v$  and  $s$  denote two kinds of residuals, both owing their existence to physical phenomena that are not well understood. The *statistically independent residual*  $v$  (for the definition of statistical independence, see §3.4)—predictable only in the statistical sense—often originates within the measuring apparatus but is not confined to it. The mean value of this residual is assumed to be zero. The *statistically dependent*

*residual*  $s$  may be considered to have originated outside the measuring system and thus can be imagined to be related to a special behaviour of the observable in a particular milieu. Even after the appropriate modelling of the effect of the environment on an observable (e.g., refraction correction for EDM distances), there remains a residual error due to an unaccounted for disturbance. The residual  $s$  is considered random with mean value zero; compared with  $v$ , however,  $s$  is statistically dependent, i.e., there is a covariance between any two elements (cf. §3.4).

The statistically dependent residual is sometimes called the signal (e.g., MORITZ [1972]). According to information theory [GOLDMAN, 1953], a quantity can be considered as *noise* if it is not wanted or as a *signal* if it contains useful information and thus is wanted. Because it is, therefore, a subjective decision whether a quantity is called a noise or a signal, other names are used here. In explaining the nature of  $s$ , it is convenient to begin by regarding the series of values  $s(\tau_i)$  as being a discrete sample of a *dynamic process* in time  $\tau$ . A dynamic process generally can be described by the following differential equation:

$$\dot{s}(\tau) = F(\tau)s(\tau), \quad (10.21)$$

where  $F$  is a function, and the dot denotes the derivative of  $s$  with respect to  $\tau$ . The solution of this equation is given by the following expression [LIEBELT, 1967]:

$$s(\tau) = S(\tau, \tau_0)s(\tau_0), \quad (10.22)$$

where  $\tau_0$  is the initial time of the process. The kernel (see §3.2)  $S$  of this equation is called the *transition function* of the process. An equivalent equation, namely,

$$s(\tau) = S(\tau, \tau_0)s(\tau_0), \quad (10.23)$$

can be written for a vectorial dynamic process  $s$ . The only difference is that, in this case, the transition function  $S$  becomes a *transition matrix*  $S$ . It is customary to talk about  $s$  as a *state vector* in time or space; these concepts will be applied in §14.6. It has been shown by various scholars that the transition matrix has the following properties [LIEBELT, 1967; PARTHASARATHY AND SCHMIDT, 1972]:

- (a)  $\dot{S}(\tau, \tau) = F(\tau).$
  - (b)  $S(\tau_1, \tau_1) = I.$
  - (c)  $S(\tau_1, \tau_2) = S^{-1}(\tau_2, \tau_1).$
  - (d)  $S(\tau_1, \tau_2)S(\tau_2, \tau_3) = S(\tau_1, \tau_3).$
  - (e)  $S(\tau_1, \tau_2)$  is always regular.
- (10.24)

Corresponding properties are equally applicable to the transition function it being a matrix of one element.

Returning now to eqn. (22), let us examine its discrete form, i.e.,

$$s(\tau_i) = S(\tau_i, \tau_j)s(\tau_j), \quad (10.25)$$

which is often called a *Markov chain* [REVUZ, 1975]. This equation can be transformed into a form containing variances and covariances. In so doing, the concept of the operator  $E$ , as introduced in §3.4, is used. (Note: this concept carries with it some probabilistic connotations which are not of interest to us yet. They will be, however, in Chapter 13, where statistical interpretations are explored; here  $E$  is simply employed as a mathematical mechanism.) Multiplying the above equation by  $s(\tau_j)$  and taking the mathematical expectation of the product yields

$$E[s(\tau_i)s(\tau_j)] = E[S(\tau_i, \tau_j)s(\tau_j)s(\tau_j)]. \quad (10.26)$$

Using the definitions of the covariance and variance as introduced in §3.4, one obtains

$$\boxed{\sigma_{s_i s_j} = S(\tau_i, \tau_j)\sigma_s^2.} \quad (10.27)$$

Evidently, if the variance of  $s$ , denoted by  $\sigma_s^2$ , is known and is constant, and if the covariance is taken as known, then the transition function can be computed from

$$S(\tau_i, \tau_j) = \sigma_{s_i s_j} \sigma_s^{-2}. \quad (10.28)$$

In this context,  $S(\tau_i, \tau_j)$  is known as the *transition probability function*; its value is always smaller than or equal to one (cf. FIG. 2). In other words, the transition function may be regarded as the probability of transition from state  $s(\tau_j)$  to state  $s(\tau_i)$ .

According to (22),  $s$  is predictable if its transition probability function (28) is known. It is useful to transform (28) further. If, for simplicity, we take homogeneity and isotropy (see §3.2), in which case we get  $S(\Delta\tau_{ij}) = S(\Delta\tau_{kl}) = S(\Delta\tau)$ , where  $\Delta\tau = \Delta\tau_{kl} = \tau_l - \tau_k = \Delta\tau_{ij} = \tau_j - \tau_i$ , (note that isotropic  $S$  does not satisfy property (24(c))) then we can write

$$S(\Delta\tau) = \sigma_{ij} \sigma_s^{-2}. \quad (10.29)$$

This represents a particularly convenient form of the transition probability function

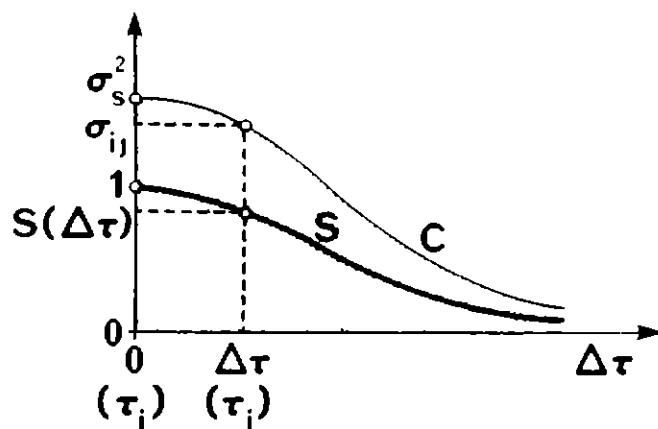


FIG. 10.2. Covariance function  $C(\Delta\tau)$  and transition probability function  $S(\Delta\tau)$ .

shown in FIG. 2. Perceived in this form, the covariance  $\sigma_{ij}$  becomes a function of  $\Delta\tau$  and, as such, is known as an *autocovariance function* (homogeneous and isotropic) which we shall be calling here simply a *covariance function*. It is usually denoted by  $C$ , so that (29) becomes

$$C(\Delta\tau) = S(\Delta\tau)\sigma_s^2. \quad (10.30)$$

Clearly, the covariance function standardized by  $\sigma_s^2$  is simply a transition probability function (cf. FIG. 2). Note that here  $\tau$  or  $\Delta\tau$  is understood to be a single parameter (e.g., time), though more complex situations may arise.

A very important concept related to this probabilistic covariance function is that concerning its limiting cases. FIG. 3 shows: the case of *total statistical dependence* (constant covariance function), which depicts the totally predictable behaviour of systematic effects; and the case of statistical independence, characterized by

$$C(\Delta\tau) = \begin{cases} C(0) \neq 0, & \Delta\tau = 0, \\ 0, & \Delta\tau \neq 0. \end{cases} \quad (10.31)$$

The covariance matrix  $C_s$  of the statistically dependent residual  $s$  (see §3.4) is fully populated. In the assumption made above, the variances were all equal, i.e.,  $\sigma_{s_1}^2 = \sigma_{s_2}^2 = \dots = \sigma_{s_N}^2 = \sigma_s^2$ . Generally, this may not always be the case and the approach described in §10.4 may have to be used. The off-diagonal elements—covariances—of the above matrix can be obtained in two ways: either directly from  $C$ , or from the transition probability function  $S(\Delta\tau)$  which, upon multiplication by  $\sigma_s^2$ , yields  $\sigma_{ij}$  (cf. eqn. (30)). The covariance matrix  $C_v$  of the statistically independent residual  $v$  is a more special case than  $C_s$ . There, the off-diagonal elements do not exist, since  $\text{cov}[v(\tau_i), v(\tau_j)] = 0$ , and  $C_v = \text{diag}(\sigma_{v_i}^2)$ , with  $\sigma_{v_i}$  all being equal. If they are not, the approach of §10.4 has to be used.

Let us now introduce a couple of conventions. First, it will be assumed that the expected values of both  $r$  and  $v$  are zero. Dropping  $\tau$ , for convenience, we get

$$E(r) = E(v + s) = E(v) + E(s) = 0, \quad (10.32)$$

and

$$E(s) = 0. \quad (10.33)$$

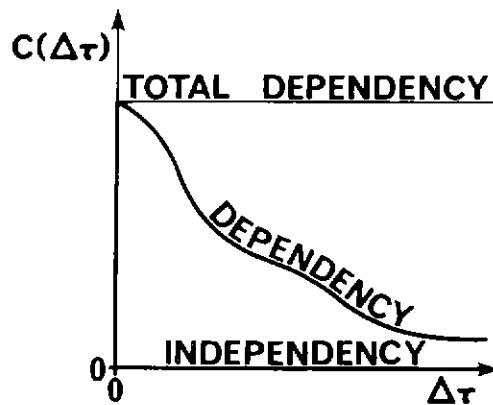


FIG. 10.3. Limiting cases of the covariance function.

Equivalently,

$$E(t - l) = E(\hat{l} + p - l) = 0, \quad (10.34)$$

as shown in FIG. 4. Further, let there be no statistical dependence between  $v$  and  $s$ . Using these conventions, the variance for residuals  $r$  is computed as follows (cf. §3.4):

$$\sigma_r^2 = E(r^2) = E[(v + s)^2] = E(v^2) + E(s^2) = \sigma_v^2 + \sigma_s^2. \quad (10.35)$$

Here, again, the probabilistic connotation of  $E$  is not invoked.

It should be pointed out that since the stochastical content of  $l$  is all within the context of  $r$ , then the uncertainty in  $r$ , of which  $\sigma_r^2$  is a measure, is also the measure of the uncertainty of  $l$ . Thus, it makes sense to define the variance of  $l$ , i.e.,  $\sigma_l^2$ , as being equal to  $\sigma_r^2$ .

In summary, the complete *decomposition of the observable* into four components is as follows:

$$l(\tau) = t(\tau) - r(\tau) = \hat{l}(\tau) + p(\tau) - v(\tau) - s(\tau). \quad (10.36)$$

It should be pointed out that the successful separation of the above components can

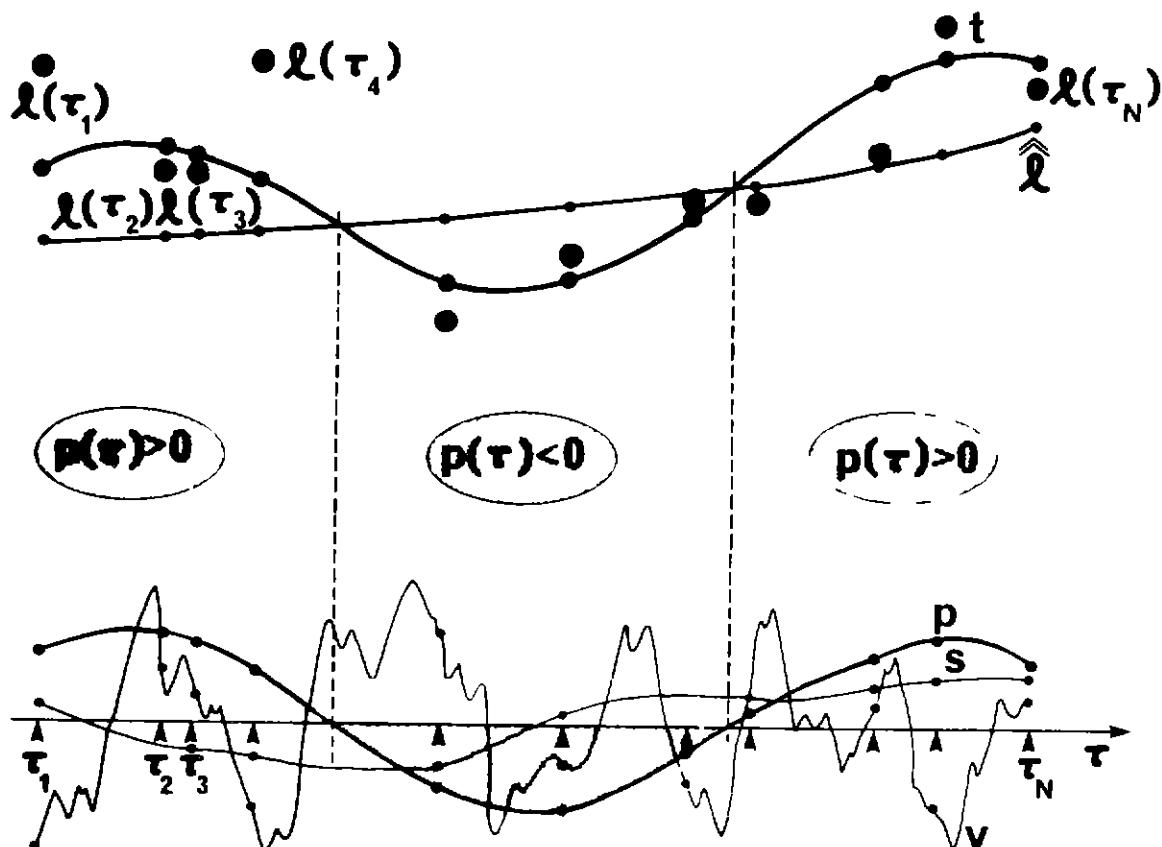


FIG. 10.4. Composition of an observable. Observation  $l(\tau)$ ; expected value of the observable  $\hat{l}(\tau)$ ; systematic component  $p(\tau)$ ; statistically independent residual  $v(\tau)$ ,  $E(v)=0$ ; statistically dependent residual  $s(\tau)$ ,  $E(s)=0$ ; also  $E(t - l) = 0$ .

only be made if there is enough information available about each. At times, it is expedient to lump various components together; this is especially the case with  $v$  and  $s$ . Note that an exact decomposition of a measured series  $\mathbf{l}(\tau)$  is impossible because of the stochastic nature of some of the components. Thus the *exact (true) value of the observable* is never known and will not be even mentioned further. It can, however, be estimated using procedures described below. Such an *estimated value of the observable* will be denoted by  $\hat{\mathbf{l}}$ .

#### 10.4. Vector of observables

Let us use the occasion of switching from observations of one observable to observations of several observables for the reconciliation of the two seemingly conflicting ideas that so far have been introduced—that of decomposing an observable with that of mathematical models. On the one hand, generalization of the decomposition formula (36) results in

$$\mathbf{l} = \hat{\mathbf{l}} + \mathbf{p} - \mathbf{v} - \mathbf{s}. \quad (10.37)$$

On the other hand, taking the model (8) and introducing the expected value of the observable results in

$$\hat{\mathbf{l}} = \mathbf{h}(\mathbf{x}), \quad (10.38)$$

or, in the linear form,

$$\hat{\mathbf{l}} = \mathbf{Hx} + \mathbf{w}. \quad (10.39)$$

Similarly, the systematic component in several dimensions reads:

$$\mathbf{p} = \Phi^T(\mathcal{T})\boldsymbol{\lambda}. \quad (10.40)$$

Substituting (40) for  $\mathbf{p}$  and (39) for  $\hat{\mathbf{l}}$  in (37) results in

$$\mathbf{l} = \mathbf{Hx} + \Phi^T(\mathcal{T})\boldsymbol{\lambda} - \mathbf{v} - \mathbf{s} + \mathbf{w}. \quad (10.41)$$

This represents the complete decomposition of the observable vector  $\mathbf{l}$  in terms of two *parameter vectors*  $\mathbf{x}$  and  $\boldsymbol{\lambda}$ , two *residual vectors*  $\mathbf{v}$  and  $\mathbf{s}$ , and the constant vector  $\mathbf{w}$ .

It is important to understand the difference between  $\mathbf{H}$  and  $\Phi(\mathcal{T})$  in the above. The design matrix  $\mathbf{H}$  is intrinsically connected with the unknown parameters being sought, thus depicting the physical and geometrical aspects underlying the experiment or project. On the other hand, the base functions of  $\Phi(\mathcal{T})$  (see §14.2) are normally selected to adequately characterize such systematic behaviour of  $\mathbf{l}$  that is of no direct interest in connection with  $\mathbf{x}$ . The latter is exemplified in EDM distances

by the unaccounted for systematic part of refraction (the statistically dependent random part being modelled by  $s$ ).

Let us now turn to the properties of the residual vectors  $v$ ,  $s$ , and  $r$ . As before,

$$E(v) = E(s) = \mathbf{0},$$

thus

$$E(r) = E(v + s) = \mathbf{0}, \quad (10.42)$$

because  $v$  and  $s$  are, again, considered mutually independent. Each component of the residual vector corresponds to a particular observable having its own physical meaning. Recall that in the univariate case (row of (16)), it was the repeated measurement of the same physical observable that produced  $N$  observations, functions of  $\tau$ . In the multivariate case, i.e., dealing with a vector of observables, only one value is dealt with for each  $l_j$ , and thus only one value of the residual is obtained for each of the  $n$   $l_j$ . These residuals may be functions of the location and the nature of the observables but, typically, not of time.

The components of  $v$  are, again, considered to be statistically independent. Thus the covariance matrix  $C_v$  of  $v$  is, again, diagonal; as before, the covariances are taken as zero. However, the variances of this matrix now have a different meaning. This time the variance is not a measure of dispersion in the sense of repeated measurements of the same quantity, but a measure of goodness of representation of the series of observations (14) by the representative value  $l_i$ . For this reason, in general,  $\sigma_{v_1}^2 \neq \sigma_{v_2}^2 \neq \dots \neq \sigma_{v_n}^2$ .

The statistically dependent residual vector  $s$  is a straightforward generalization of the statistically dependent residual  $s$ . A parallel development will not be made for this more general situation, but the main consequences of the generalization will simply be stated. First, the covariance function  $C(\Delta\tau)$  may have a different scale for each component  $s_i$  of  $s$ ; thus, in general,  $\sigma_{s_1}^2 \neq \sigma_{s_2}^2 \neq \dots \neq \sigma_{s_n}^2$ . The meaning attached to  $\Delta\tau$  is now typically that of the space coordinate, i.e., distance; it is stressed, however, that  $\Delta\tau$  may be a more complicated entity than just distance. For two components  $s_i, s_j$  with different variances  $\sigma_i^2, \sigma_j^2$  their covariance  $\sigma_{ij}$  can be evaluated from the following equation:

$$\sigma_{ij} = C_{ij}(\Delta\tau) = \sigma_i \sigma_j \text{cov}(\Delta\tau), \quad (10.43)$$

where  $C_{ij}$  is the *cross-covariance function* of  $s_i$  and  $s_j$  [LIEBELT, 1967]. Then the complete covariance matrix  $C_s$  can be assembled.

To conclude, clearly, two alternatives of the decomposition of an observable exist: The first is to consider the whole table of observables  $L$  (16) together, and solve the problem of decomposition in one step. The second is to treat each series  $l_j(\tau_i)$  (14) separately, and then construct the vector  $l$  (15) to be used in the main mathematical model. The latter is the normal practice. In the second option, the first step is known as *data series analysis*, treated in §14.2, by which it is possible to examine separately the behaviour of each observable in time. The objective of this preprocessing is to

learn something, for instance, about a series of measured distances of one line or sea level readings at a single tide gauge before introducing them; along with other quantities, into the main model. The outcome of the data series analysis will be one value  $t_i$  for each observable along with its variance and any covariances that may exist. Thus the covariance matrix of  $t$ ,

$$\mathbf{C}_t = \mathbf{C}_v + \mathbf{C}_s, \quad (10.44)$$

should be available as a result of the preprocessing.

## CHAPTER 11

# CLASSES OF MATHEMATICAL MODELS

In this chapter, mathematical models are first classified, and then each class is discussed. The first section contains a classification scheme for all the models encountered in geodesy. Also introduced in this section is the idea of measuring the quality of solution, metric and Hilbert spaces. The second section discusses the simplest class of models those that have a unique solution, together with the transformation of uncertainties—the covariance law. An overview of underdetermined models is given in the third section, along with one particular prescription for how to deal with them. The final section establishes the foundation for the treatment of overdetermined models. However, the most widely used technique for treating this most frequently encountered class of models is left to Chapter 12.

### 11.1. Classification of models

The possible forms of a mathematical model were discussed in §10.2. Our discussion of models will now be continued, this time from the point of view of their solution, with the type of solution being the basis for model classification. The definition of the solution, in a narrow sense, is the determination of the vector of parameters  $\mathbf{x}$  or, possibly,  $\lambda$ . Because geodesists are also interested in the accuracy of  $\mathbf{x}$ , the covariance matrix  $\mathbf{C}_x$  is sought as well. Thus the solution, in a more general sense, is the pair of quantities  $(\mathbf{x}, \mathbf{C}_x)$ . The act of solving a mathematical model is then equivalent to seeking the transformation

$$(\mathbf{l}, \mathbf{C}_l) \rightarrow (\mathbf{x}, \mathbf{C}_x). \quad (11.1)$$

The context will indicate which of the two senses of the solution is intended.

There are three possible classes of solutions to linear or linearized models: unique, underdetermined, and overdetermined. Models that cannot be linearized usually require different techniques for their solution. As shown in FIG. 1, the treatment of a model is based on its form, linearizeability, and determinacy. First the model is checked to see if it is explicit in  $\mathbf{x}$ ; if it is, the model has a *unique solution* since  $\mathbf{x}$  is evaluated directly from the model. The only complication then is in the determination of  $\mathbf{C}_x$ , which will be dealt with in detail in §11.2.

If the model is not explicit in  $\mathbf{x}$  and is not naturally linear, it has to be linearized. The linearizeability of the model, which, at this stage, could be either explicit in  $\mathbf{l}$  or

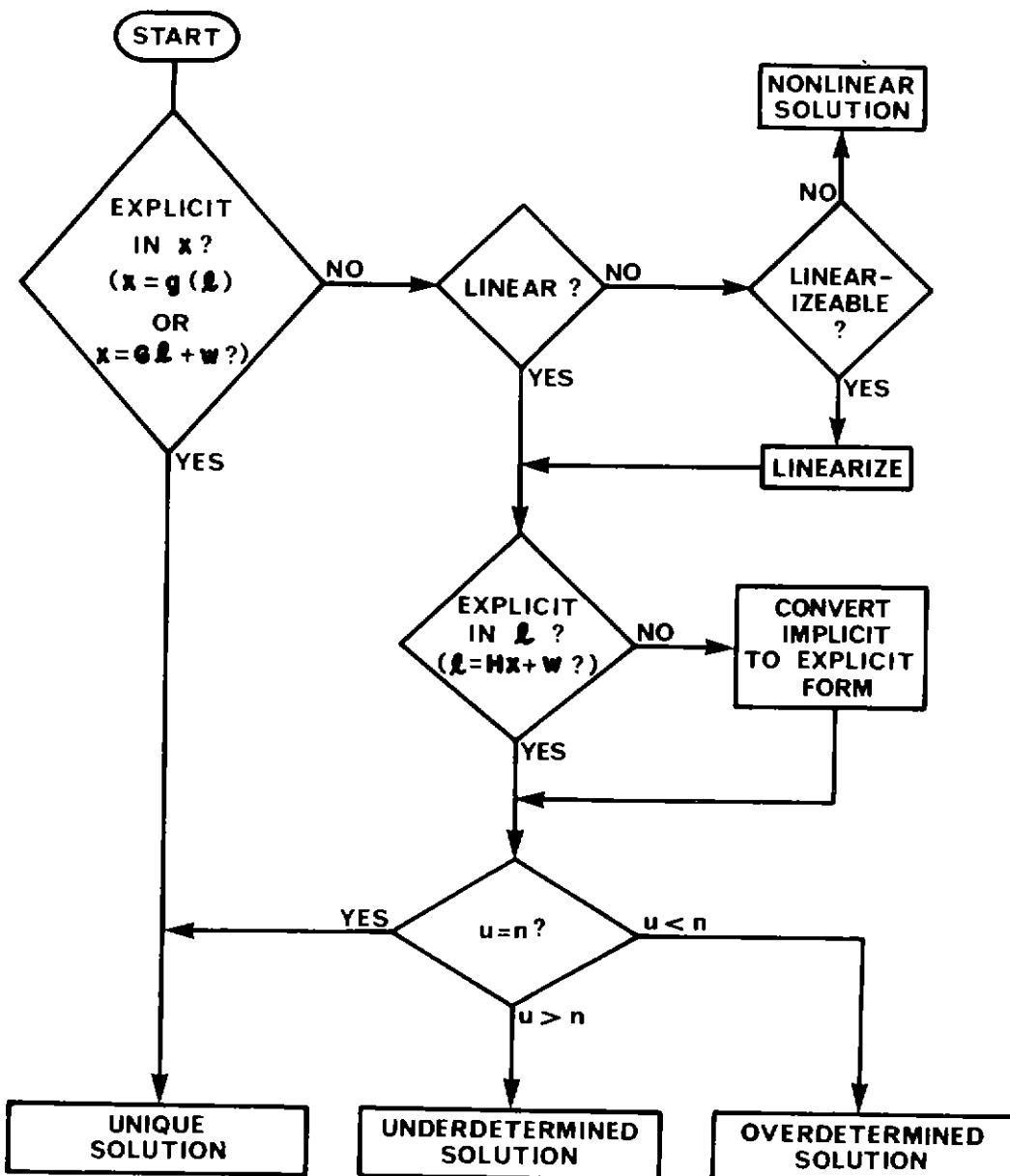


FIG. 11.1. Treatment of models. (Here,  $u$  and  $n$  are respectively the number of unknown parameters and observations.)

implicit, is examined next. If the model cannot be linearized, then the answer must be sought in the domain of *non-linear solutions*. Models in this class are characterized by unknown parameters that are locked into functional expressions and cannot be brought to a linear form through linearization. Special mathematical techniques—such as spectral analysis (see §14.2)—are needed to solve these models, and no special effort is made here to explain these.

To linearize the more general implicit form  $f(x, l) = \mathbf{0}$ , the model is approximated with a multidimensional linear Taylor's series (see §3.2); namely,

$$f(x, l) \doteq f(x^{(0)}, l^{(0)}) + \frac{\partial f}{\partial x} \Big|_{x=x^{(0)}, l=l^{(0)}} (x - x^{(0)}) + \frac{\partial f}{\partial l} \Big|_{x=x^{(0)}, l=l^{(0)}} (l - l^{(0)}) \doteq \mathbf{0}. \quad (11.2)$$

This equation is identical in form to (10.10) when we put

$$\mathbf{A} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \Big|_{\begin{array}{l} \mathbf{x} = \mathbf{x}^{(0)}, \\ \mathbf{l} = \mathbf{l}^{(0)} \end{array}}, \quad \mathbf{B} = \frac{\partial \mathbf{f}}{\partial \mathbf{l}} \Big|_{\begin{array}{l} \mathbf{x} = \mathbf{x}^{(0)}, \\ \mathbf{l} = \mathbf{l}^{(0)} \end{array}}, \quad (11.3)$$

and

$$\mathbf{w} \doteq -\mathbf{Ax}^{(0)} - \mathbf{Bl}^{(0)} + \mathbf{f}(\mathbf{x}^{(0)}, \mathbf{l}^{(0)}). \quad (11.4)$$

Here  $\mathbf{x}^{(0)}$  is the *point of expansion* in the parameter space  $\mathcal{X}$ , and  $\mathbf{l}^{(0)}$  is the point of expansion in the observation space  $\mathcal{L}$ . The matrices  $\mathbf{A}$  and  $\mathbf{B}$  can be regarded as Jacobian matrices of transformation (see §3.1) from parameter and observation spaces to the model space  $\mathcal{T}$ , valid for a small neighbourhood of  $\mathbf{x}^{(0)}$  and  $\mathbf{l}^{(0)}$ . For reasonably well-behaved functions  $\mathbf{f}$ , the solution of the linearized model can be regarded as valid within the neighbourhood of the two points  $\mathbf{x}^{(0)}$  and  $\mathbf{l}^{(0)}$ . Since higher order terms in (4) are neglected, iterations such as those described in §12.1 may be required to get a better approximation of the solution.

Following the same approach for the model explicit in  $\mathbf{l}$  (10.8), an equation identical to (10.9) is arrived at where

$$\mathbf{H} = \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \Big|_{\mathbf{x} = \mathbf{x}^{(0)}} \quad \text{and} \quad \mathbf{w} = -\mathbf{Hx}^{(0)} + \mathbf{l}^{(0)}. \quad (11.5)$$

Note that the term  $\mathbf{Hx}^{(0)}$  can have the meaning of either  $\mathbf{Ax}$  or  $\Phi^T(\mathcal{T})\lambda$  or both combined, as explained in §10.4. Also, clearly, the linear model explicit in  $\mathbf{l}$  is only a special case of the linear implicit model, where  $\mathbf{H} = \mathbf{A}$  and  $-\mathbf{I} = \mathbf{B}$ .

Assuming now that the model is linear, either naturally or through linearization, one can inquire if the model is explicit (in  $\mathbf{l}$ ) or implicit. An implicit model can be converted into an explicit model simply by defining new *quasi-observations*  $\tilde{\mathbf{l}}$  as a linear combination of the observations  $\mathbf{l}$ :

$$\tilde{\mathbf{l}} = -\mathbf{Bl}. \quad (11.6)$$

Substitution of this in (10.11) yields

$$\tilde{\mathbf{l}} = \mathbf{Ax} + \mathbf{w}, \quad (11.7)$$

which is a model explicit in  $\tilde{\mathbf{l}}$  and can be treated in the same way as (10.9). Sometimes, however, it is expedient to work with the implicit form itself, because it might afford a deeper insight into the nature of the observation vector  $\mathbf{l}$ , as will be seen in Chapter 12.

At this juncture, an examination of the possible uniqueness of the solution is appropriate. If  $u = n$  and  $\mathbf{H}$  is regular (see §3.1), then the model has a unique solution; namely,

$$\mathbf{x} = \mathbf{H}^{-1}(\mathbf{l} - \mathbf{w}). \quad (11.8)$$

If, however,  $\mathbf{H}$  is singular there is no unique solution and generalized inverses need be employed, as will be shown in §14.5. If  $u < n$ , one is faced with an *overdetermined solution*  $\mathbf{x}$ ; this case is treated in §11.4. If  $u > n$ , we speak of *underdetermined solutions*; one such solution is described in §11.3.

Within the class of models with underdetermined solutions, two particular situations may arise:  $\text{rank } \mathbf{H} = n < u$  or  $\text{rank } \mathbf{H} < n < u$ . Within the class of models with overdetermined solutions, the same two situations are possible. Thus it is apparent that the problem of rank deficiency ( $\text{rank } \mathbf{H} < \min \dim \mathbf{H}$ ) exists in all three classes of models. A detailed treatment of this problem will be given in §11.4 and §14.5 for the overdetermined class only. While mathematical models for  $u = n$  clearly pose no uniqueness problem when  $\mathbf{H}$  has full rank, this is not so for the models with underdetermined solutions. There are an infinite number of solutions  $\mathbf{x}$  exist, that satisfy the equations even when  $\mathbf{H}$  has full rank. All that can be done is to solve for some of the unknown parameters as functions of other unknown parameters. When the solution is overdetermined, there is generally no  $\mathbf{x}$  that would satisfy the system of equations, and special steps must be taken to cure the situation (see §11.4).

Being faced with an infinite number of solutions, one should first examine the possibilities of selecting the best solution. The choice of the *best solution* is always somewhat arbitrary, but meaningful results can be obtained if some general rules are followed. These rules were formulated in mathematics within the theory of metric spaces; the concepts will be briefly explained here. Let us begin by assuming a space (finite or infinite) in which it is possible to measure a distance  $\rho(a, b)$  between any two of its elements  $a, b$ . For any space, this distance, or *metric*, can be selected in many ways. To be meaningful, however, any such distance  $\rho(a, b)$  must satisfy the following relations:

- (a)  $\rho(a, b)$  must be a non-negative real number ( $\rho(a, b) = 0$  if and only if  $a = b$ );
- (b)  $\rho(a, b) = \rho(b, a)$ ; and
- (c)  $\rho(a, b) \leq \rho(a, c) + \rho(c, b)$ .

These relations are known as *axioms of a metric*.

A space in which a specific metric is defined is called a *metric space*. If the space happens to be already *normed*, i.e., if there is the operation called norm (usually symbolized by  $\|a\|$ ; note that one way of defining norm has already been shown in §3.2) defined in the space, then, because of its properties, the norm can also be used as a metric [DAVIS, 1963; CHENEY, 1966]. The norm is akin to the metric through the following relation:

$$\|a\| = \rho(a, 0),$$

and

$$\|a - b\| = \rho(a, b), \quad (11.9)$$

where 0 is the null element. In the following, both the distance and the norm are used.

To select the best of the infinite number of solutions of the mathematical model, first the space in which to measure the 'goodness' of the solution must be selected, then the metric with which to measure the goodness must be defined. In other words, a metric space must be chosen which in turn implies the best solution. Once the choice of the appropriate metric space has been made, the selection of the best solution then requires the imposition of the *minimum distance condition* or *minimum norm condition*.

The most commonly used metric is the *Euclidean metric*—the one used in ordinary Euclidean geometrical space. In more general contexts, the Euclidian distance (for the orthonormal coordinate basis) is defined as

$$\rho(\mathbf{a}, \mathbf{b}) = \left[ \sum_{i=1}^n (a_i - b_i)^2 \right]^{1/2},$$

where it is assumed that the space is a linear vector space of dimension  $n$  [COTLAR AND CIGNOLI, 1974], and thus  $\mathbf{a}$  and  $\mathbf{b}$  are vectors. This assumption is not detrimental to the argument here because, as seen already, in geodesy, vectors ( $\mathbf{x}, \mathbf{l}, \mathbf{r}$ , etc.) are always dealt with. Note that since  $\mathbf{a}, \mathbf{b}$  are vectors (to be denoted as  $\mathbf{a}, \mathbf{b}$ ), the above equation can be written in matrix notation as the square root of the scalar product of two vectors, i.e.,

$$\rho(\mathbf{a}, \mathbf{b}) = [(\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{b})]^{1/2}. \quad (11.10)$$

It can easily be verified that this distance, also called the *mean quadratic distance*, satisfies the axioms of a metric. In the literature, the mean quadratic distance is often replaced by the *quadratic norm*.

Other examples of a metric often used in mathematics are

(a) uniform (Tchebychev's) metric

$$\rho(\mathbf{a}, \mathbf{b}) = \max_{i=1, \dots, n} |a_i - b_i|, \quad (11.11)$$

(b)  $q$ -metric

$$\rho(\mathbf{a}, \mathbf{b}) = \left( \sum_{i=1}^n |a_i - b_i|^q \right)^{1/q}. \quad (11.12)$$

It has been shown that the least-squares and uniform metrics respectively are special cases of  $q$ -metric for  $q=2$  and  $q \rightarrow \infty$ . Cases in which  $q \neq 2$  are not commonly used in geodesy and are not dealt with here. The interested reader is referred to DAVIS [1963] and SINGER [1970].

Let us return to the Euclidean metric and interpret it geometrically. The realization that  $\mathbf{a}$  and  $\mathbf{b}$  are position vectors in  $n$ -dimensional space, with the same scale on all the  $n$  axes, allows us to construct a vector diagram of this situation in a two-dimensional Euclidean space (see FIG. 2). In a more general case, if the scales on the coordinate axes are not the same, then one speaks of an orthogonal coordinate basis. In such a case, the square of the distance becomes

$$\rho^2(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|^2 = (\mathbf{a} - \mathbf{b})^T \mathbf{P} (\mathbf{a} - \mathbf{b}),$$

(11.13)

where  $\mathbf{P}$  is a diagonal matrix. The elements of  $\mathbf{P}$  are different constants representing squares of scales on the individual axes. The right-hand side of (13) is a quadratic form (cf. §3.1). If an even more general coordinate basis is considered in the Euclidean space, then  $\mathbf{P}$  becomes more complicated; it becomes a non-diagonal, positive-definite, symmetrical matrix. Regardless of how complicated the coordinate

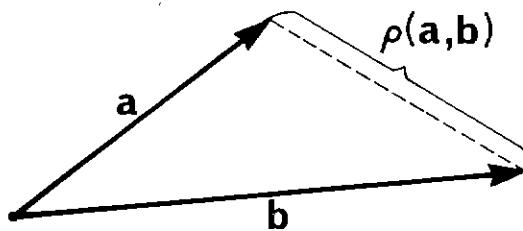


FIG. 11.2. Distance in Euclidean space.

basis is, the least-squares distance remains the square root of the scalar product ( $\mathbf{a} - \mathbf{b}$ ) with itself. The matrix  $P$  plays the role of the *metric tensor* in the space under consideration—a concept borrowed from differential geometry (e.g., SYNGE AND SCHILD [1949]; WREDE [1963]). Note that the first Euclidean form for the distance (eqn. (10)) is a special case of the above in which  $P$  equals  $I$ , i.e., when the metric tensor is equal to an identity matrix.

If the distance between two elements in a complete, Euclidean space can be written as the square root of a scalar product of the difference of these two elements with itself, i.e., if the scalar product is defined as above, then this space is called a *Hilbert space*. The mathematical theory of Hilbert spaces is very well developed, and a wealth of useful theorems can be found in the literature (e.g., LUENBERGER [1969]; SINGER [1970]). Minimization of distances between various elements of a Hilbert space is sometimes called *Hilbert space optimization*. This terminology is adopted here as well. The reasons why Hilbert space methods are particularly appropriate for geodesy are elucidated in §13.1.

It is worthwhile to point out, at this juncture, that a completely parallel structure can be built on compact spaces (with infinite dimensions), where the analogue of the least-squares norm is defined by (3.42). Spaces with this norm are called  $L_2$  spaces [COTLAR AND CIGNOLI, 1974]. It is the minimization of this kind of norm that leads to the theory of generalized Fourier series (eigenfunction developments), as mentioned in §3.2. An example of its application in geodesy will be seen in §20.2 and elsewhere.

## 11.2. Models with a unique solution

In this section, models of both the explicit and implicit forms are treated; for each form, the solution for the parameters  $x$  and the corresponding covariance matrix  $C_x$  are given.

(a) The solution  $x$  of the model explicit in  $x$  ((10.3) or (10.4)) is obtained by direct evaluation, since  $t$  is known. Derivation of the covariance matrix  $C_x$ , on the other hand, requires that the model, if not naturally linear, be linearized first. Let us begin with a model of the explicit form that is already linear (10.4) and apply the mathematical expectation  $E$  to it. This yields

$$E(x) = E(Gt + w) = GE(t) + w, \quad (11.14)$$

because  $\mathbf{G}$  is known and thus is not a variable in the expectation operation, nor is  $w$ . Subtracting (14) from (10.4) gives

$$\mathbf{x} - \mathbf{E}(\mathbf{x}) = \mathbf{G}[\mathbf{l} - \mathbf{E}(\mathbf{l})]. \quad (11.15)$$

Taking the defining eqn. (3.100) for  $\mathbf{C}_x$ , and substituting from (15), one gets

$$\mathbf{C}_x = \mathbf{E}\{\mathbf{G}[\mathbf{l} - \mathbf{E}(\mathbf{l})][\mathbf{l} - \mathbf{E}(\mathbf{l})]^T \mathbf{G}^T\}, \quad (11.16)$$

which can be rewritten as

$$\mathbf{C}_x = \mathbf{G}\mathbf{E}\{[\mathbf{l} - \mathbf{E}(\mathbf{l})][\mathbf{l} - \mathbf{E}(\mathbf{l})]^T\}\mathbf{G}^T,$$

or

$$\boxed{\mathbf{C}_x = \mathbf{G}\mathbf{C}_l\mathbf{G}^T,} \quad (11.17)$$

because the covariance matrix of the observation vector  $\mathbf{l}$  is, clearly,

$$\mathbf{C}_l = \mathbf{E}\{[\mathbf{l} - \mathbf{E}(\mathbf{l})][\mathbf{l} - \mathbf{E}(\mathbf{l})]^T\}. \quad (11.18)$$

Equation (17) is known as the *covariance law*. Uncertainties in the vector  $\mathbf{l}$ , given in terms of the covariance matrix  $\mathbf{C}_l$ , can be traced into  $\mathbf{x}$  and characterized by  $\mathbf{C}_x$ .

The special case of  $\mathbf{x} = \mathbf{g}(\mathbf{l})$ , in which  $\mathbf{x}$  has only one element  $x$  and the individual elements in the vector  $\mathbf{l}$  are statistically independent, i.e.,  $\mathbf{C}_l$  is a diagonal matrix, yields

$$x = g(l), \quad (11.19)$$

with variance

$$\sigma_x^2 = \left( \frac{\partial g}{\partial l_1} \sigma_{l_1} \right)^2 + \left( \frac{\partial g}{\partial l_2} \sigma_{l_2} \right)^2 + \cdots + \left( \frac{\partial g}{\partial l_n} \sigma_{l_n} \right)^2 = \sum_{i=1}^n \left( \frac{\partial g}{\partial l_i} \sigma_{l_i} \right)^2. \quad (11.20)$$

The above expression is known as the law of *propagation of random errors*. If the model expressed by (19) is linear, i.e.,

$$x = \mathbf{G}\mathbf{l} + w, \quad (11.21)$$

where  $\mathbf{G} = [g_1, g_2, \dots, g_n]$ , then the variance of  $x$  is given directly by

$$\sigma_x^2 = \sum_{i=1}^n (g_i \sigma_{l_i})^2. \quad (11.22)$$

If the model for  $x$  is, for instance, just the mean of the individual, statistically independent  $l$ 's acquired with equal accuracy, i.e., if

$$x = \frac{1}{n} \sum_1^n l_i, \quad (11.23)$$

then

$$\sigma_x^2 = \frac{1}{n} \sigma_l^2. \quad (11.24)$$

For the case described by (21), the *propagation of systematic errors* is given by

$$\delta x = \sum_{l=1}^n g_l \delta l_l, \quad (11.25)$$

where  $\delta x$  is the total systematic error, while  $\delta l_l$  are the individual systematic errors in the observations. This follows from the rules for total differentiation (§3.2).

(b) The solution of the model explicit in  $l$  requires, again, linearization if the model is not already linear. Equation (8) is obtained from the linearized form (10.9), where  $\mathbf{H}^{-1}$  exists because  $\mathbf{H}$  is a square, and assumed regular, matrix. The corresponding covariance matrix is derived by applying the covariance law to (8). This results in

$$\mathbf{C}_x = \mathbf{H}^{-1} \mathbf{C}_l (\mathbf{H}^{-1})^T. \quad (11.26)$$

(c) The solution of the implicit model also requires linearization. Recalling (7),

$$\mathbf{x} = \mathbf{A}^{-1}(\tilde{l} - \mathbf{w}),$$

because  $\mathbf{A}$  is square and regular. Using (6),

$$\mathbf{x} = -\mathbf{A}^{-1}(\mathbf{B}l + \mathbf{w}). \quad (11.27)$$

As the reader can verify, the corresponding covariance matrix follows again directly from the covariance law,

$$\mathbf{C}_x = \mathbf{A}^{-1} \mathbf{B} \mathbf{C}_l \mathbf{B}^T (\mathbf{A}^{-1})^T. \quad (11.28)$$

### 11.3. Models with an underdetermined solution

Models with an underdetermined solution occur when their linearized version has fewer equations than unknown parameters, i.e.,  $n < u$ . This is a result of insufficient observations for the determination of the sought after parameters and rarely occurs in geodesy. On the other hand, as we will see in §11.4, the reformulation of models with overdetermined solution leads to the underdetermined case and, as such, will be treated extensively later. For this reason, only a very partial treatment to the underdetermined case is given below.

The problem can be stated: Solve for  $\mathbf{x}$  from (10.9), which can be rewritten here as

$$\mathbf{H}\mathbf{x} = \mathbf{l} - \mathbf{w}, \quad (11.29)$$

under the conditions that  $n < u$ . Only the explicit model is treated because any implicit model can be transformed into the explicit form, as shown in §11.1. Since

the design matrix  $\mathbf{H}$  is not regular, its (regular) inverse cannot be computed. Nevertheless, there is an infinite number of solutions for  $\mathbf{x}$ : the question then narrows down to which of these solutions should one select. Only one way is going to be shown here—that of the orthogonal decomposition of  $\mathbf{H}$ —as this procedure can be used to select an  $\mathbf{x}$  such that the criterion of minimum norm of  $\mathbf{x}$  is satisfied. It makes sense to use this criterion clearly only if there are physical or mathematical reasons for expecting the  $\mathbf{x}$  vector to be small. The procedure [LAWSON AND HANSON, 1974] is summarized below:

- (a) To begin with, a vector  $\mathbf{y}$  ( $\dim \mathbf{y} = \dim \mathbf{x} = u$ ) is introduced such that

$$\mathbf{x} = \mathbf{Q} \mathbf{y}, \quad (11.30)$$

where  $\mathbf{Q}$  is an orthogonal matrix;  $\dim \mathbf{Q} = (u, u)$ .

- (b) Replacing  $\mathbf{x}$  from the above in the model (29) yields

$$\mathbf{H}\mathbf{Q}\mathbf{y} = \mathbf{l} - \mathbf{w}. \quad (11.31)$$

- (c)  $\mathbf{Q}$  is selected such that

$$\mathbf{H}\mathbf{Q} = [\mathbf{R} | \mathbf{0}], \quad (11.32)$$

where  $\mathbf{R}$  is a lower triangular, non-singular matrix;  $\dim \mathbf{R} = (n, n)$ . This selection is called the *orthogonal decomposition* of  $\mathbf{H}$  and is a unique operation.

- (d) The solution of (31) is

$$\mathbf{y}_1 = \mathbf{R}^{-1}(\mathbf{l} - \mathbf{w}).$$

The above equation yields unique values for the first  $n$  components of  $\mathbf{y}$ , denoted here by  $y_1$ , but does not give any values for the remaining  $n - u$  components, denoted by  $y_2$ . Thus the complete solution should read:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \quad (11.33)$$

in which  $y_2$  may be selected arbitrarily.

- (e) Often  $y_2$  is selected as a null vector. This selection leads to the minimization of  $\|\mathbf{y}\|$  and, subsequently, if the metric on  $\mathfrak{X}$  is Euclidean, to the minimization of  $\|\mathbf{x}\|$ . However, this selection is only meaningful if the criterion of minimum  $\|\mathbf{x}\|$  to be enforced is meaningful.

Note that the covariance matrix  $\mathbf{C}_t$  is not needed in this approach. Let us only mention here one particularly elegant attribute of the above described method. Although  $\mathbf{H}$  is singular, one can get a special kind of inverse for it, the pseudo-inverse  $\mathbf{H}^+$  (see, e.g., RAO AND MITRA [1971]; BEN-ISRAEL AND GREVILLE [1974]), so that

$$\mathbf{x} = \mathbf{H}^+(\mathbf{l} - \mathbf{w}). \quad (11.34)$$

The  $\mathbf{x}$  obtained in this way has identical properties to the above solution. Some other alternative ways of selecting a suitable solution will be given in §14.5.

#### 11.4. Models with an overdetermined solution

When the linearized model has more equations than unknown parameters, the solution is overdetermined. Thus overdetermined models indicate that more observations than necessary have been taken for the determination of the unknowns. This results in the inability to obtain any  $\mathbf{x}$ . To understand why, consider the linearized model (10.9), in which  $n > u$  and  $\text{rank } \mathbf{H} = u < n$ . Arbitrarily select  $u$  of the  $n$  equations, and get a solution for  $\mathbf{x}$ . Then choose a different subset of  $u$  equations, and get another solution for  $\mathbf{x}$ . These two solutions are generally different: mathematically, the two subsets of  $u$  equations are generally inconsistent. This dilemma continues when other subsets of  $u$  equations are considered. Models with overdetermined solutions are the ‘daily bread’ of geodesists, thus this topic will be treated in some depth here and again in Chapter 12.

As already mentioned in §11.1, the problem of overdetermination can be alleviated by *reformulation of the model*, which consists of using the expected rather than observed values of observables, i.e.,

$$\hat{\mathbf{l}} = \mathbf{l} + \mathbf{r} = \mathbf{Hx} + \mathbf{w}. \quad (11.35)$$

The unknown residual vector  $\mathbf{r}$  is introduced to make the equations consistent by allowing the observations to change (from  $\mathbf{l}$  to  $\hat{\mathbf{l}} + \mathbf{r}$ ). Note that the reformulation of the model in terms of the expected value  $\hat{\mathbf{l}}$  of the observable agrees with the decomposition scheme of (10.37), when  $\mathbf{p} = \mathbf{Hx}$  is assumed. Therefore, (35) is equivalent to (10.41) for  $\lambda = \mathbf{0}$  and  $\mathbf{v} + \mathbf{s} = \mathbf{r}$ . This formulation leads to an infinite number of solutions, because now both  $\mathbf{x}$  and  $\mathbf{r}$  are unknown vectors. Prior to reformulation, the situation was characterized by a redundancy of  $n - u$ . After reformulation, the redundancy is decreased by  $n$ , because  $n$  new quantities were added to be solved for, leaving fewer equations than unknowns; i.e.,  $(n - u) - n < 0$ . Under these circumstances, the best that can be done is to solve for certain unknown parameters as a function of other unknown parameters. Thus we find ourselves in the situation of underdetermined solutions once again.

The best solution is normally sought in the observation space. The metric selected is usually the least-squares distance, so that  $\mathcal{L}$  becomes a metric space with the least-squares metric. The outstanding properties of this metric space are that the solution is comparatively easy from the mathematical point of view, as will be seen in Chapter 12, and the quantities involved lend themselves to an easy statistical interpretation, as will be seen in Chapter 13. The *least-squares solution* is obtained through the minimization of the distance between  $\mathbf{r}$  and  $\mathbf{0}$  or, equivalently (cf. (35)), between  $\mathbf{Hx}$  and  $\mathbf{l} - \mathbf{w}$ . This can be written as

$$\min_{\mathbf{x} \in \mathcal{X}, \mathbf{r} \in \mathcal{L}} \rho(\mathbf{Hx}, \mathbf{l} - \mathbf{w}),$$

or, more briefly, as

$$\min_{\mathbf{x}, \mathbf{r}} \rho(\mathbf{r}, \mathbf{0}) = \min_{\mathbf{x}, \mathbf{r}} \|\mathbf{r}\|. \quad (11.36)$$

Note that since the problem has been reformulated, there are now two unknown

vectors ( $\mathbf{x}$  and  $\mathbf{r}$ ), and the minimization must be carried out with respect to both. The  $\mathcal{L}$  is now an affine Euclidean space, and its metric tensor is normally selected to be equal to the inverse of the covariance matrix  $\mathbf{C}_l$  of the vector  $\mathbf{l}$ . The rationale for this selection is based on statistical considerations and will be treated in §13.1. Equation (36) can be written as

$$\min_{\mathbf{x}, \mathbf{r}} [(\mathbf{Hx} - \mathbf{l} + \mathbf{w})^T \mathbf{C}_l^{-1} (\mathbf{Hx} - \mathbf{l} + \mathbf{w})] = \min_{\mathbf{x}, \mathbf{r}} \mathbf{r}^T \mathbf{C}_r^{-1} \mathbf{r}, \quad (11.37)$$

where it has already been shown (eqn. (10.44)) that  $\mathbf{C}_l \equiv \mathbf{C}_r$ . Note that the minimum distance condition actually involves minimization, in the above sense, of the residuals  $\mathbf{r}$  of the observations. This is why, in the literature, the above condition is often referred to as the *minimum quadratic form of the weighted residuals*, or as the *minimum sum of squares of the weighted residuals*. The latter designation is true only when  $\mathbf{s}$  is missing in  $\mathbf{r}$  so that  $\mathbf{r} = \mathbf{v}$ , i.e., when the matrix  $\mathbf{C}_r$  is diagonal. The various least-squares techniques found in the literature are merely variations on the theme of the minimum condition given by (37). These techniques are obtained by stipulating different vectors and metric tensors in the quadratic form, as will be shown later in the appropriate places. Generally, whichever least-squares technique is used, it can always be regarded as a variety of Hilbert space optimization, which is why the philosophy of the Hilbert space approach was discussed in §11.1.

One may ask: Why cannot models with underdetermined solutions be treated the same way? They indeed can! The difference, however, is that while there is a clear justification for the requirement that  $\mathbf{r}$  should be minimized, one way or another, one can seldom find a similar justification for minimizing the solved for parameters. This is why the two classes of models are usually treated by very different techniques.

In the mathematical models, it has been tacitly assumed that the observables are of a physical or geometrical nature. This is not, however, the only possibility. In designing the model, other aspects of the experiment, such as logistics and cost, may be taken into account. These considerations could alter the perception of the observables, but not the mathematical concepts described above. When such *economic parameters* are involved, it is at times desirable to impose a priori constraints on the solution, as discussed in §10.2. When these constraints are formulated as inequalities, rather than equalities, the realm of linear programming or quadratic programming is involved. For a discussion of these disciplines, the interested reader is referred to, e.g., HADLEY [1964].

## CHAPTER 12

### **LEAST-SQUARES SOLUTION OF OVERDETERMINED MODELS**

In this chapter, an overview of the least-squares approach to solving over-determined models is given, the main concepts are explained, and the most important equations are developed. In the first section, the formulation of the least-squares problem for overdetermined models is discussed. Once the problem is formulated, the solution is then given in the second section. At the outset of this section, some alternative solutions are briefly discussed, while the remainder deals with the standard Lagrange method of finding the solution under the least-squares condition. The last section contains the derivation of the covariance matrices for the results. Throughout this chapter, the implicit mathematical model is utilized, because it is the most general. The reader interested in an in-depth treatment of the least-squares technique is referred to, e.g., HIRVONEN [1971], BJERHAMMAR [1973], and MIKHAIL [1976].

#### **12.1. Formulation of the least-squares problem**

As shown in §11.4, overdetermined models have generally inconsistent equations, and reformulation of the model is necessary to remove these inconsistencies. Accordingly (10.10) becomes

$$\mathbf{f}(\mathbf{x}, \hat{\mathbf{l}}) = \mathbf{f}(\mathbf{x}, \mathbf{l} + \hat{\mathbf{r}}) = \mathbf{0}, \quad (12.1)$$

where  $\hat{\mathbf{r}}$  are the *expected residuals*. Since no confusion can ensue here, in this chapter we shall use  $\mathbf{r}$  for  $\hat{\mathbf{r}}$ . To simplify the solution, the above model is normally approximated with the linear part of a Taylor series, whether the model is linear or not, in a fashion analogous to the linearization discussed in §11.1. For the points of expansion, usually the observed values of observables and approximate values ( $\mathbf{x}^{(0)}$ ) of the unknown parameters are chosen resulting in (cf. (11.2))

$$\begin{aligned} \mathbf{f}(\mathbf{x}, \mathbf{l}) &= \mathbf{f}(\mathbf{x}^{(0)} + \boldsymbol{\delta}, \mathbf{l}^{(0)} + \mathbf{r}) \\ &\doteq \mathbf{f}(\mathbf{x}^{(0)}, \mathbf{l}^{(0)}) + \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \Big|_{\mathbf{l}=\mathbf{l}^{(0)}} (\mathbf{x} - \mathbf{x}^{(0)}) + \frac{\partial \mathbf{f}}{\partial \mathbf{l}} \Big|_{\mathbf{x}=\mathbf{x}^{(0)}} (\mathbf{l} - \mathbf{l}^{(0)}) \doteq \mathbf{0}, \end{aligned}$$

or simply

$$\boxed{\mathbf{A}\delta + \mathbf{B}r + w = \mathbf{0}.} \quad (12.2)$$

We observe that this equation is merely the differential form of the original non-linear mathematical model and describes the relation of the quantities in the neighbourhoods of  $x^{(0)}$ ,  $\ell^{(0)}$ , and  $w$ —see FIG. 1. The design matrices  $\mathbf{A}$ ,  $\mathbf{B}$  are given by (11.3); their presence here is understandable since the differential form of the model is formulated in  $m$ -dimensional model space  $\mathcal{F}$ , and thus all the other quantities must be transformed to  $\mathcal{F}$ . This is verified simply by noting that each of the terms in (2)—namely,  $\mathbf{A}\delta$ ,  $\mathbf{B}r$ , and  $w$ —are  $m$ -dimensional vectors. The quantities  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $w$  are known, while  $\delta$  and  $r$  are unknown. The vector of corrections  $\delta$  to the approximate parameters  $x^{(0)}$  is clearly a special kind of solution vector, and the constant vector

$$w = f(x^{(0)}, \ell^{(0)}) \quad (12.3)$$

is, in this context, called the *misclosure vector*. In this chapter, no specific meaning is placed on the residual vector  $r$  in terms of further decomposition into parts. In §14.3,  $r$  will be decomposed into  $v$  and  $s$  as the need arises, while here it is regarded only as a single vector.

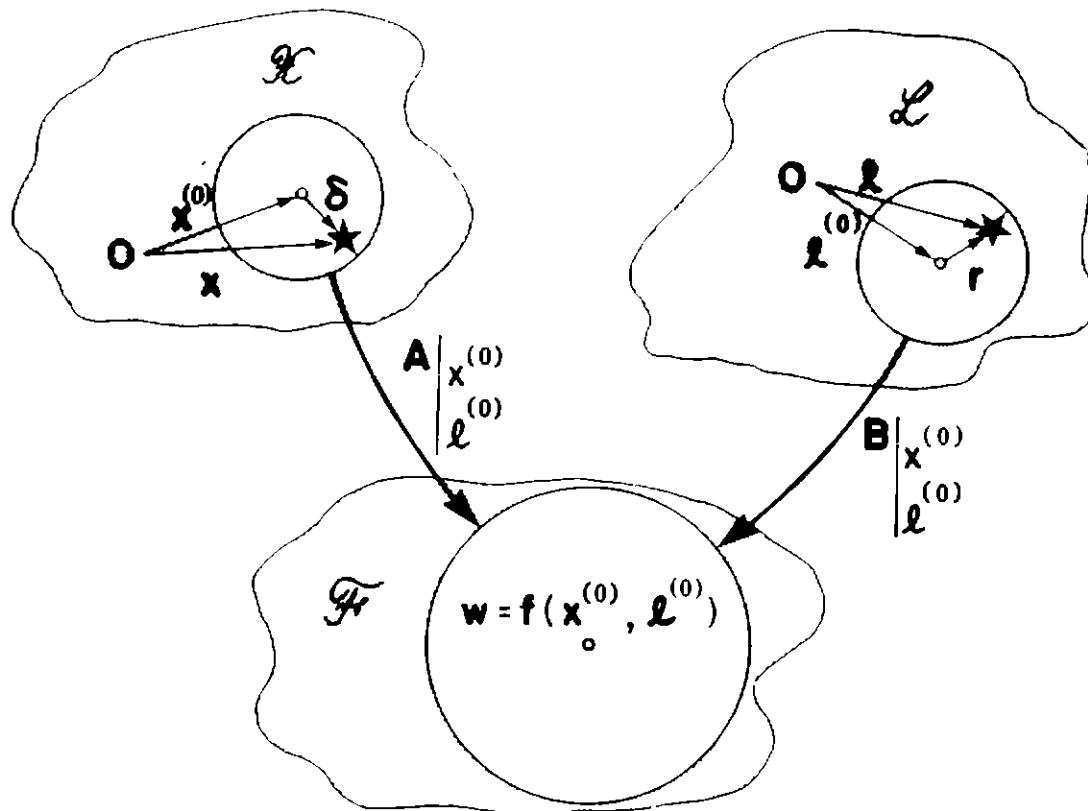


FIG. 12.1. Linearization.

To solve for  $\delta$ , either the length of  $r$  can be minimized in  $\mathcal{L}$  or the length of its projection,

$$\tilde{r} = -Br \quad (12.4)$$

(cf. (11.6)), can be minimized in  $\mathcal{F}$ . The first case leads to

$$\min_{\delta, r} (r^T C_r^{-1} r), \quad (12.5)$$

while the second case corresponds to

$$\min_{\delta, \tilde{r}} (\tilde{r}^T C_{\tilde{r}}^{-1} \tilde{r}). \quad (12.6)$$

Using the second approach, the linearized model becomes

$$\tilde{r} = A\delta + w. \quad (12.7)$$

The covariance matrix  $C_{\tilde{r}}$ , which is needed to metricize the model space, is induced by  $r$  and thus follows from the application of the covariance law to (4): namely,

$$C_{\tilde{r}} = BC_r B^T = M^{-1}. \quad (12.8)$$

Notice that the formulation in  $\mathcal{F}$  allows for the rewriting of the minimum condition directly in terms of  $\delta$ . Substitution for  $\tilde{r}$  in (6) from (7) yields

$$\min_{\delta, \tilde{r}} \tilde{r}^T C_{\tilde{r}}^{-1} \tilde{r} = \min_{\delta, r} [(A\delta + w)^T C_r^{-1} (A\delta + w)]. \quad (12.9)$$

Formulation in  $\mathcal{L}$  does not allow such a direct substitution and thus has to be treated differently. It will be shown in the next section that, indeed, both formulations lead to the same results.

## 12.2. Solution of the least-squares problem

The minimization of the quadratic form, given by (9), is an extremum problem of mathematics. The standard method of finding an extremum is described in §3.2 and here only the result is given; namely,

$$\frac{\partial}{\partial \tilde{r}} (\tilde{r}^T C_{\tilde{r}}^{-1} \tilde{r}) = \frac{\partial}{\partial \delta} [(A\delta + w)^T C_r^{-1} (A\delta + w)] = 0.$$

Carrying out the differentiation results in (see §3.2)

$$C_{\tilde{r}}^{-1} \hat{r} = (A^T C_r^{-1} A) \delta + A^T C_r^{-1} w = 0. \quad (12.10)$$

Substituting for  $\hat{r}$  from (7) into the leftmost expression and premultiplying it by  $A^T$  gives an equation identical to the second equation. Clearly, if one of the distances (quadratic forms) is minimized then so is the other, and it does not matter which one is used in achieving this. Equation (10) is a matrix equation called the *system of*

(linear least-squares) *normal equations*, where  $\mathbf{A}^T \mathbf{C}_r^{-1} \mathbf{A}$  is called the *coefficient matrix*, and  $\mathbf{A}^T \mathbf{C}_r^{-1} \mathbf{w}$  is a constant vector.

In order to determine whether the extremum is the maximum or minimum, we have to take the second derivative of one of the quadratic forms. It equals to  $\mathbf{A}^T \mathbf{C}_r^{-1} \mathbf{A}$ , which is a positive-definite matrix because in  $\mathbf{C}_r = \mathbf{B} \mathbf{C}_r \mathbf{B}^T$ ,  $\mathbf{C}_r$  is, by definition, a positive-definite matrix, and  $\mathbf{A}$  is of rank  $u \leq n$  (cf. §3.1). Thus the extremum is a minimum, which is what is wanted.

In (10), a cap has been used over  $\delta$  to distinguish it from  $\delta$  without a cap in, e.g., (2). The reason is that the imposition of the minimum quadratic distance defines the *least-squares solution*  $\hat{\delta}$ —sometimes called the *vector of corrections*—which is unique among the countless acceptable solutions  $\delta$ . Similarly,

$$\hat{\mathbf{x}} = \mathbf{x}^{(0)} + \hat{\delta} \quad (12.11)$$

is called the *least-squares estimate of the unknown parameters*  $\mathbf{x}$ , or simply the *adjusted parameters*.

Let us now try to find the solution  $\hat{\delta}$  by minimizing the length of the residual vector in observation space instead of model space, i.e., by using the condition expressed by (5) instead of (6). As pointed out in §11.1, a solution in observation space has the advantage of yielding directly the values of the residuals  $\mathbf{r}$  rather than  $\tilde{\mathbf{r}}$ . These  $\tilde{\mathbf{r}}$  cannot always be transformed into  $\mathbf{r}$  because, generally,  $\mathbf{B}$  is not a regular matrix. Since a direct substitution for  $\mathbf{r}$  in (5) from (2) is not possible, another approach is needed. The approach offered by Lagrange [KORN AND KORN, 1968], which is normally used, is based on the idea of involving the other quantities from the model through the following mathematical trick. The vector equation (2) is multiplied by an arbitrary vector  $\mathbf{k}$  from  $\mathcal{F}$ . The resulting scalar product is zero for any finite  $\mathbf{k}$ —because the right-hand side is a null vector—and can thus be added to the quadratic form  $(\mathbf{r}^T \mathbf{C}_r^{-1} \mathbf{r})$  without affecting its value. Hence, the following expression is valid:

$$\mathbf{r}^T \mathbf{C}_r^{-1} \mathbf{r} = \mathbf{r}^T \mathbf{C}_r^{-1} \mathbf{r} + 2\mathbf{k}^T (\mathbf{A}\delta + \mathbf{B}\mathbf{r} + \mathbf{w}) = \phi. \quad (12.12)$$

In the literature, the expression to be minimized, i.e., the sum  $\phi$  of the two scalar products, is called the *variation function*. The vector of the *Lagrange correlates*  $\mathbf{k}$ , being from  $\mathcal{F}$ , has a dimension  $m$ ; it is unspecified to begin with and, thus, plays the same role as the unknown vectors  $\delta$  and  $\mathbf{r}$ .

It has been shown [HANCOCK, 1917; HADLEY, 1964] that the addition of the scalar product involving the correlates to the variation function does not change the location of the extremum. The normal equations are again obtained by taking the derivatives of  $\phi$  with respect to  $\delta$ ,  $\mathbf{r}$ , and  $\mathbf{k}$ , and equating them to null vectors. The result is (cf. §3.2)

$$\frac{1}{2} \frac{\partial \phi}{\partial \mathbf{r}} = \tilde{\mathbf{r}}^T \mathbf{C}_r^{-1} + \hat{\mathbf{k}}^T \mathbf{B} = \mathbf{0}, \quad (12.13)$$

$$\frac{1}{2} \frac{\partial \phi}{\partial \delta} = \hat{\mathbf{k}}^T \mathbf{A} = \mathbf{0}, \quad (12.14)$$

and

$$\frac{1}{2} \frac{\partial \phi}{\partial \boldsymbol{k}} = \mathbf{A}\hat{\boldsymbol{\delta}} + \mathbf{B}\hat{\boldsymbol{r}} + \mathbf{w} = \mathbf{0}. \quad (12.15)$$

The transpose of (13), (14), together with (15) constitute the desired system of normal equations; namely,

$$\mathbf{C}_r^{-1}\hat{\boldsymbol{r}} + \mathbf{B}^T\hat{\boldsymbol{k}} = \mathbf{0}, \quad (12.16)$$

$$\mathbf{A}^T\hat{\boldsymbol{k}} = \mathbf{0}, \quad (12.17)$$

$$\mathbf{A}\hat{\boldsymbol{\delta}} + \mathbf{B}\hat{\boldsymbol{r}} + \mathbf{w} = \mathbf{0}. \quad (12.18)$$

Note that the overhead cap is again inserted to show that these are the least-squares estimates of the unknown vectors that are being dealt with. The normal equation system in hypermatrix form is written as (see §3.1)

$$\begin{bmatrix} \mathbf{C}_r^{-1} & \mathbf{B}^T & \mathbf{0} \\ \mathbf{B} & \mathbf{0} & \mathbf{A} \\ \mathbf{0} & \mathbf{A}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{r}} \\ \hat{\boldsymbol{k}} \\ \hat{\boldsymbol{\delta}} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{w} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad (12.19)$$

with a coefficient matrix of dimensions  $n + m + u$ . The hypervector, comprised of  $\hat{\boldsymbol{r}}$ ,  $\hat{\boldsymbol{k}}$ , and  $\hat{\boldsymbol{\delta}}$ , can be solved for by inverting the coefficient matrix directly. This is uneconomical, however, because of its size and the presence of null matrices. Thus, a normal equation system is derived below in which the matrix inversions are smaller and the null matrices suppressed.

To begin with, the residual vector  $\hat{\boldsymbol{r}}$  is eliminated from the system using the technique of partitioning the solution (see §3.1). The result is

$$\begin{bmatrix} -\mathbf{M}^{-1} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{k}} \\ \hat{\boldsymbol{\delta}} \end{bmatrix} + \begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \quad (12.20)$$

Next,  $\hat{\boldsymbol{k}}$  is eliminated by using the same technique to yield

$$(\mathbf{A}^T \mathbf{M} \mathbf{A})\hat{\boldsymbol{\delta}} + \mathbf{A}^T \mathbf{M} \mathbf{w} = \mathbf{0}. \quad (12.21)$$

This is the resulting system of normal equations deduced by the *Lagrange method*. Comparison of the above with (10) shows that minimization in  $\mathcal{L}$  yields an identical result  $\hat{\boldsymbol{\delta}}$  to that obtained by minimization in  $\mathcal{F}$ .

The least-squares solution is then written as

$$\hat{\boldsymbol{\delta}} = -(\mathbf{A}^T \mathbf{M} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{M} \mathbf{w}. \quad (12.22)$$

Note that the solution is unique only if the rank of the coefficient matrix,

$$\mathbf{N} = \mathbf{A}^T \mathbf{M} \mathbf{A} = \mathbf{A}^T (\mathbf{B} \mathbf{C}, \mathbf{B}^T)^{-1} \mathbf{A}, \quad (12.23)$$

is equal to  $u$ . If it is lower, generalized matrix inversions must be used, as will be

shown in §14.5. For the sake of brevity, the following notation is introduced:

$$\mathbf{u} = \mathbf{A}^T \mathbf{M} \mathbf{w}, \quad (12.24)$$

where  $\mathbf{M}$  is clearly the weight matrix of observations transformed into  $\mathcal{F}$ . In this notation, the least-squares normal equations are

$$\mathbf{N} \hat{\boldsymbol{\delta}} + \mathbf{u} = \mathbf{0}, \quad (12.25)$$

and the solution is

$$\hat{\boldsymbol{\delta}} = -\mathbf{N}^{-1} \mathbf{u}. \quad (12.26)$$

The correlates  $\hat{\mathbf{k}}$  are solved for by using the first equation from (20),

$$-\mathbf{M}^{-1} \hat{\mathbf{k}} + \mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w} = \mathbf{0}, \quad (12.27)$$

thus,

$$\hat{\mathbf{k}} = \mathbf{M}(\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}). \quad (12.28)$$

The *least-squares residuals*  $\hat{\mathbf{r}}$  are arrived at by using (16), which yields

$$\hat{\mathbf{r}} = -\mathbf{C}_r \mathbf{B}^T \hat{\mathbf{k}} = -\mathbf{C}_r \mathbf{B}^T \mathbf{M}(\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}). \quad (12.29)$$

Finally, the *least-squares estimate of the observations* is

$$\hat{\mathbf{l}} = \mathbf{l} + \hat{\mathbf{r}}. \quad (12.30)$$

A discussion on how to handle the so-called *effect of non-linearity*, alluded to in §11.1, is now in order. It arises from having replaced a non-linear model by its linear approximation. When analysing this problem, POPE [1974] employed the Newton–Gauss iterative method for solving non-linear problems and discovered that the non-linear solution can be arrived at by a series of repeated linear solutions. The linearized model at the  $n$ th iteration is given by

$$\mathbf{A}^{(n)}(\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}) + \mathbf{B}^{(n)}(\mathbf{l}^{(n+1)} - \mathbf{l}^{(n)}) + \mathbf{f}(\mathbf{x}^{(n)}, \mathbf{l}^{(n)}) = \mathbf{0}, \quad (12.31)$$

where  $(\mathbf{x}^{(n)}, \mathbf{l}^{(n)})$  is the latest point of expansion; and  $\mathbf{A}^{(n)}$ ,  $\mathbf{B}^{(n)}$  are evaluated at this point of expansion—see FIG. 2. The solutions for  $\boldsymbol{\delta}^{(n)} = \mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}$  and  $\mathbf{r}^{(n)} = \mathbf{l}^{(n+1)} - \mathbf{l}^{(n)}$  are still obtained using (26) and (30) but this time with  $\mathbf{w}^{(n)} = \mathbf{f}(\mathbf{x}^{(n)}, \mathbf{l}^{(n)}) + \mathbf{B}^{(n)}(\mathbf{l}^{(0)} - \mathbf{l}^{(n)})$ , where  $\mathbf{l}^{(0)}$  is the vector of observed values. Note the presence of the second term that makes the  $n$ th iteration different from the initial solution, i.e., if  $\mathbf{l}^{(n)} = \mathbf{l} = \mathbf{l}^{(0)}$  is dealt with, then the expression for  $\mathbf{w}^{(n)} = \mathbf{w}^{(0)}$  reduces to eqn. (3). According to POPE [1974], the iterations should be carried out until two successive increments, i.e.,  $\boldsymbol{\delta}^{(n)}$  and  $\boldsymbol{\delta}^{(n+1)}$ , go to zero.

Updating both the parameters and observations is peculiar to the implicit non-linear model because the model is non-linear in both quantities. Models explicit in  $\mathbf{l}$

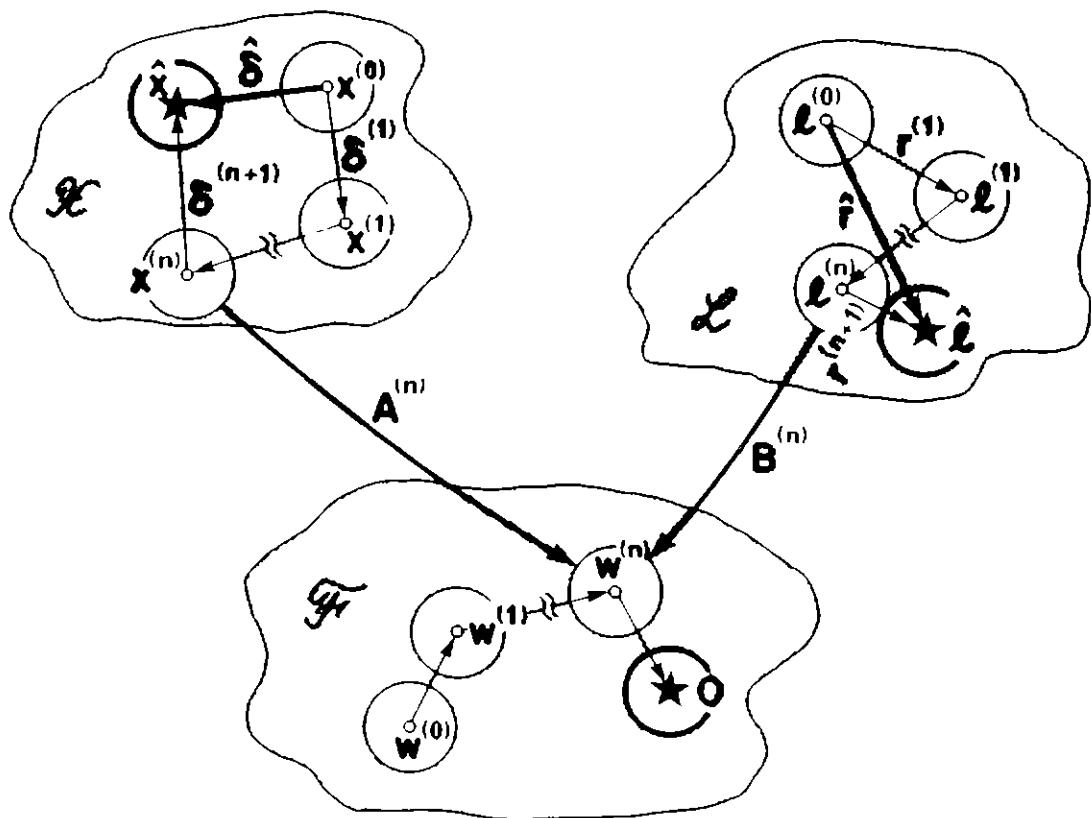


FIG. 12.2. Iterations of a linearized implicit model.

need be iterated only in  $x$  since they are already linear in  $\ell$ . The iterations of the model explicit in  $\ell$  are carried out as long as the magnitude of increments in  $x$ , i.e.,  $\delta^{(n)}$ , remains significant [MIKHAIL, 1976]. Condition models need be iterated in  $\ell$ , the same way as the parametric models are iterated in  $x$ .

The aforementioned approach to solving the least-squares problem, based on the use of normal equations, is only one of several possibilities. There are other computational methods which do not employ normal equations but work with the design matrix  $A$  itself. These include *Householder's orthogonal transformation* and *singular value analysis* (e.g., LAWSON AND HANSON [1974]). The main difference between the methods using normal equations and those working directly with the design matrix is that the former methods require only about half as many operations as do the latter. However, when using the normal equations approach, a computational precision is required that is equal to the square of that needed when using the Householder algorithm. The normal equations approach is used herein partly because of the advantage mentioned above, but chiefly because this is the approach most often used in geodesy. The reason why it is preferred in geodesy is that it is more versatile when it comes to deducing the covariance matrices of the results.

It also has to be pointed out that even within the normal equation approach there are more options than those presented here. Thus, for instance, one does not have to linearize the (non-linear) mathematical model but simply get the *normal equations in*

*non-linear form.* These are simply the partial derivatives of the non-linear model, i.e.,

$$\frac{\partial \mathbf{f}}{\partial \mathbf{r}} = \mathbf{0}, \quad (12.32)$$

and

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \mathbf{0}. \quad (12.33)$$

The solution of this system of non-linear, simultaneous equations can then be attempted using any of a number of techniques (see, e.g., BEREZIN AND ZHIDKOV [1962]) including the above-mentioned Newton-Gauss.

For solving the linear least-squares normal equations, the *method of Cholesky* [THOMPSON, 1969] is most widely used because it is designed expressly to handle the positive-definite, symmetrical matrices found in the systems of least-squares normal equations (e.g., KNIGHT AND STEEVES [1974], HANSON [1976]). Various other techniques are used in practice when either very large systems or systems with sparse matrices are concerned. More about these will be found in §18.2.

### 12.3. Covariance matrices of the results

Having obtained the solutions for the unknown parameters and the residuals, we are now faced with the problem of deriving the covariance matrices characterizing their accuracies. These covariance matrices are indispensable to the statistical evaluation of the results, as will be seen in Chapter 13. In this section, the covariance matrices for the misclosure vector  $\mathbf{w}$ , the corrections  $\hat{\delta}$ , the residuals  $\hat{\mathbf{r}}$ , and the adjusted observations  $\hat{\mathbf{l}}$  are derived.

(a) The misclosure vector is defined by (3). Application of the covariance law to this equation yields

$$\boxed{\mathbf{C}_w = \mathbf{B}\mathbf{C}_r\mathbf{B}^T = \mathbf{M}^{-1}}, \quad (12.34)$$

where  $\mathbf{B} = \partial \mathbf{f} / \partial \mathbf{l}$  is the usual second design matrix defined in §11.1, and  $\mathbf{C}_r = \mathbf{C}_l$ . It is interesting to note that  $\mathbf{C}_w = \mathbf{C}_{\hat{\mathbf{r}}}$  as one can discover by comparing (34) to (8). This matrix can be determined before the least-squares problem is solved and has any meaning only if the  $\mathbf{w}$  can be regarded as a statistically meaningful quantity. This question will be examined more closely, from the statistical point of view, in §13.4.

(b) The least-squares estimate of the parameters ((26) and (11)) can be written as

$$\hat{\mathbf{x}} = \mathbf{x}^{(0)} - \mathbf{N}^{-1}\mathbf{A}^T\mathbf{M}\mathbf{w}. \quad (12.35)$$

Applying the covariance law to the above equation yields

$$\mathbf{C}_{\hat{\mathbf{x}}} = (-\mathbf{N}^{-1}\mathbf{A}^T\mathbf{M})\mathbf{M}^{-1}(-\mathbf{N}^{-1}\mathbf{A}^T\mathbf{M})^T,$$

and after carrying out the prescribed operations, one gets

$$\mathbf{C}_{\hat{x}} = \mathbf{N}^{-1} = (\mathbf{A}^T \mathbf{M} \mathbf{A})^{-1} = \mathbf{C}_{\hat{\delta}}. \quad (12.36)$$

The covariance matrix of  $\hat{\delta}$  is identical with that of  $\hat{x}$  since  $\mathbf{x}^{(0)}$  is a constant vector and thus its covariance matrix is equal to a null matrix. In a way, the derivation of  $\mathbf{C}_{\hat{x}}$  completes the transformation, given by (11.1), we set out to seek. There are, however, other quantities of interest that can also now be derived.

(c) Replacing  $\hat{\delta}$  in (29) from (26) yields

$$\hat{r} = \mathbf{C}_r \mathbf{B}^T (\mathbf{M} \mathbf{A} \mathbf{N}^{-1} \mathbf{A}^T \mathbf{M} - \mathbf{M}) \mathbf{w} = -\mathbf{C}_r \mathbf{B}^T \mathbf{L} \mathbf{w}, \quad (12.37)$$

which is an expression involving only one independent variable  $\mathbf{w}$ . It is interesting to note that  $\mathbf{L}$  is the matrix linking the misclosure vector with the correlates, i.e.,  $\hat{k} = \mathbf{L} \mathbf{w}$ . Applying the covariance law to (37) results in

$$\mathbf{C}_{\hat{r}} = \mathbf{C}_r \mathbf{B}^T \mathbf{L} \mathbf{C}_w (\mathbf{C}_r \mathbf{B}^T \mathbf{L})^T.$$

Substitution for  $\mathbf{C}_w$  and completion of the prescribed operations gives

$$\mathbf{C}_{\hat{r}} = \mathbf{C}_r \mathbf{B}^T \mathbf{L} \mathbf{B} \mathbf{C}_r = \mathbf{C}_r \mathbf{B}^T \mathbf{M} [\mathbf{I} - \mathbf{A} (\mathbf{A}^T \mathbf{M} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{M}] \mathbf{B} \mathbf{C}_r. \quad (12.38)$$

This covariance matrix is always singular and plays an important role in the rejection of erroneous observations; it will be treated in detail in §13.4.

(d) The expression for the adjusted observations (30) can also be written as

$$\hat{l} = \mathbf{l} - \mathbf{C}_r \mathbf{B}^T \mathbf{M} (\mathbf{A} \hat{\delta} + \mathbf{w}).$$

Replacing  $\hat{\delta}$  from (26) and  $\mathbf{w}$  from (3) yields

$$\hat{l} = \mathbf{l} - \mathbf{C}_r \mathbf{B}^T \mathbf{L} f(\mathbf{x}^{(0)}, \mathbf{l}), \quad (12.39)$$

which is an expression containing only one variable  $\mathbf{l}$ . Applying the covariance law to the above yields

$$\mathbf{C}_{\hat{l}} = \left( \frac{\partial \hat{l}}{\partial \mathbf{l}} \right) \mathbf{C}_l \left( \frac{\partial \hat{l}}{\partial \mathbf{l}} \right)^T,$$

where

$$\frac{\partial \hat{l}}{\partial \mathbf{l}} = \mathbf{l} - \mathbf{C}_r \mathbf{B}^T \mathbf{L} \mathbf{B}. \quad (12.40)$$

Substitution results in

$$\mathbf{C}_{\hat{l}} = \mathbf{C}_l - \mathbf{C}_r \mathbf{B}^T \mathbf{L} \mathbf{B} \mathbf{C}_r.$$

Comparison of the above covariance matrix with that of  $\hat{\mathbf{C}}$  shows that

$$\hat{\mathbf{C}}_t = \mathbf{C}_t - \hat{\mathbf{C}}_r = \mathbf{C}_r - \hat{\mathbf{C}}_r; \quad (12.41)$$

as expected, the variances of the adjusted observations are smaller than the variances before adjustment.

It is also possible to develop the corresponding *cross-covariance matrices* that describe the amount of statistical dependence between pairs of vectors. For instance,  $\hat{\mathbf{C}}_{\hat{\mathbf{r}}\hat{\mathbf{s}}}$  describes the cross-covariance between  $\hat{\mathbf{r}}$  and  $\hat{\mathbf{s}}$ . Since there is no direct use for these matrices, they will not be derived here. A limited use of cross-covariance matrices will be made, however, in §14.3, where the pertinent expressions will be derived.

It should be pointed out that the above covariance matrices are exact only for linear models. For non-linear models, it is necessary to account for the effect of non-linearity. This results in more complicated expressions involving *Hessian matrices*, i.e., matrices of second derivatives [CELMINS, 1973; POPE, 1974]. In most cases, however, the linear expressions found here give a realistic enough indication of the accuracy.

Practical application of the least-squares technique often runs into the problem of what to do if the scale of the covariance matrix of the observations is not known, i.e., if only the relative size of the elements of  $\mathbf{C}_t$  is known. To resolve this problem, let us denote the inverse of the matrix of such arbitrarily scaled elements by  $\mathbf{P}_t$ . Then the following relation holds:

$$\mathbf{P}_t = \sigma_0^2 \mathbf{C}_t^{-1} = \sigma_0^2 \mathbf{C}_r^{-1}, \quad (12.42)$$

where the scale factor  $\sigma_0^2$  is referred to as the *variance factor*, and the matrix  $\mathbf{P}_t$  is called the *weight matrix*. Let us now investigate the consequences of not knowing  $\sigma_0^2$  before the adjustment is carried out.

An investigation of eqn. (26) leads to the discovery that an a priori knowledge of  $\sigma_0^2$  is unnecessary to arrive at the correct value for  $\hat{\mathbf{d}}$  because, clearly,

$$\hat{\mathbf{d}} = -(\mathbf{A}^T \mathbf{M} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{M} \mathbf{w} = -[\mathbf{A}^T (\mathbf{B} \mathbf{P}_t^{-1} \mathbf{B}^T)^{-1} \mathbf{A}]^{-1} \mathbf{A}^T (\mathbf{B} \mathbf{P}_t^{-1} \mathbf{B}^T)^{-1} \mathbf{w}. \quad (12.43)$$

Similarly, from (29),

$$\hat{\mathbf{r}} = -\mathbf{C}_r \mathbf{B}^T \mathbf{M} (\mathbf{A} \hat{\mathbf{d}} + \mathbf{w}) = -\mathbf{P}_t^{-1} \mathbf{B}^T (\mathbf{B} \mathbf{P}_t^{-1} \mathbf{B}^T)^{-1} (\mathbf{A} \hat{\mathbf{d}} + \mathbf{w}), \quad (12.44)$$

which shows that  $\hat{\mathbf{r}}$  is not dependent upon the a priori knowledge of  $\sigma_0^2$  either. All four covariance matrices,  $\mathbf{C}_w$ ,  $\mathbf{C}_{\delta}$ ,  $\mathbf{C}_r$ , and  $\hat{\mathbf{C}}_t$ , however, are dependent upon  $\sigma_0^2$  and thus directly affected by the lack of knowledge of the scale. The conclusion is that a value for the scale factor  $\sigma_0^2$  must be known at least before the covariance matrices can be evaluated. Fortunately, even when  $\sigma_0^2$  is not known beforehand, its value can be estimated from the results.

Before deriving the estimated, or a posteriori, value of  $\sigma_0^2$ , three intermediate steps have to be taken. Firstly, from the normal equations (25) becomes

$$\mathbf{A}^T \mathbf{M} (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}) = \mathbf{0},$$

and, upon transposition,

$$(\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w})^T \mathbf{M} \mathbf{A} = \mathbf{0}. \quad (12.45)$$

Secondly, the above equation can also be written as

$$\mathbf{A}^T \mathbf{M} \mathbf{w} = -\mathbf{A}^T \mathbf{M} \mathbf{A} \hat{\boldsymbol{\delta}}. \quad (12.46)$$

Thirdly, the quadratic form of the residual (eqn. (12)) can be transformed by employing (28) and (29) for  $\hat{\mathbf{k}}$  and  $\hat{\mathbf{r}}$  as follows:

$$\begin{aligned} \hat{\mathbf{r}}^T \mathbf{C}_r^{-1} \hat{\mathbf{r}} &= (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w})^T \mathbf{M} \mathbf{B} \mathbf{C}_r \mathbf{C}_r^{-1} \mathbf{C}_r \mathbf{B}^T \mathbf{M} (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}) \\ &= (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w})^T \mathbf{M} \mathbf{A} \hat{\boldsymbol{\delta}} + (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w})^T \mathbf{M} \mathbf{w}. \end{aligned}$$

After utilizing eqns. (45) and (46) in the above,

$$\hat{\mathbf{r}}^T \mathbf{C}_r^{-1} \hat{\mathbf{r}} = -\hat{\boldsymbol{\delta}}^T \mathbf{A}^T \mathbf{M} \mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}^T \mathbf{M} \mathbf{w} = -\hat{\boldsymbol{\delta}}^T \mathbf{N} \hat{\boldsymbol{\delta}} + \mathbf{w}^T \mathbf{M} \mathbf{w}. \quad (12.47)$$

With the realization that  $\mathbf{w}^T \mathbf{M} \mathbf{w} = \text{tr}(\mathbf{w} \mathbf{w}^T \mathbf{M})$ —see §3.1—the expected value of the quadratic form is obtained by taking the mathematical expectation of each of the two constituent quadratic forms. To begin with, one has

$$E(\mathbf{w}^T \mathbf{M} \mathbf{w}) = E[\text{tr}(\mathbf{w} \mathbf{w}^T \mathbf{M})] = \text{tr}[E(\mathbf{w} \mathbf{w}^T) \mathbf{M}].$$

By definition,

$$E\{[\mathbf{w} - E(\mathbf{w})][\mathbf{w} - E(\mathbf{w})]^T\} = E(\mathbf{w} \mathbf{w}^T) - E(\mathbf{w}) E(\mathbf{w}^T) = \mathbf{C}_w = \mathbf{M}^{-1},$$

so that

$$E(\mathbf{w} \mathbf{w}^T) = \mathbf{M}^{-1} + E(\mathbf{w}) E(\mathbf{w}^T).$$

Thus, the expected value of the above quadratic form is

$$E(\mathbf{w}^T \mathbf{M} \mathbf{w}) = \text{tr}(\mathbf{M}^{-1} \mathbf{M}) + \text{tr}[E(\mathbf{w}) E(\mathbf{w}^T) \mathbf{M}] = m + c, \quad (12.48)$$

where  $c$  is a real number. Similarly,

$$\begin{aligned} E(\hat{\boldsymbol{\delta}}^T \mathbf{N} \hat{\boldsymbol{\delta}}) &= E[\text{tr}(\hat{\boldsymbol{\delta}} \hat{\boldsymbol{\delta}}^T \mathbf{N})] = \text{tr}[E(\hat{\boldsymbol{\delta}} \hat{\boldsymbol{\delta}}^T) \mathbf{N}] \\ &= \text{tr}(\mathbf{N}^{-1} \mathbf{N}) + \text{tr}[E(\hat{\boldsymbol{\delta}}) E(\hat{\boldsymbol{\delta}}^T) \mathbf{N}] = u + d. \end{aligned} \quad (12.49)$$

It has been shown that  $c = d$  [BLAHA, 1978], thus leaving

$$E(\hat{\mathbf{r}}^T \mathbf{C}_r^{-1} \hat{\mathbf{r}}) = m - u. \quad (12.50)$$

Let us now rewrite the above in terms of the weight matrix  $\mathbf{P}_l$  using eqn. (42): namely,

$$\mathbb{E}\left(\frac{1}{\sigma_0^2} \hat{\mathbf{r}}^T \mathbf{P}_l \hat{\mathbf{r}}\right) = m - u.$$

After carrying out the mathematical expectation, treating  $\hat{\mathbf{r}}$  as known, and denoting  $\mathbb{E}(\sigma_0^2)$  by  $\hat{\sigma}_0^2$ , the final result for the *a posteriori variance factor* is

$$\hat{\sigma}_0^2 = \frac{\hat{\mathbf{r}}^T \mathbf{P}_l \hat{\mathbf{r}}}{m - u},$$

(12.51)

where

$$m - u = v \quad (12.52)$$

defines the number of *degrees of freedom*, also called *redundancy*.

Once the variance factor is estimated, the covariance matrices of the results can also be estimated. To begin with, let us rewrite the covariance matrix for the misclosure vector  $\mathbf{w}$ . Use of (34) and (42) results in

$$\mathbf{C}_w = \mathbf{B} \sigma_0^2 \mathbf{P}_l^{-1} \mathbf{B}^T = \sigma_0^2 \tilde{\mathbf{M}}^{-1}, \quad (12.53)$$

where  $\tilde{\mathbf{M}}^{-1}$  is the improperly scaled covariance matrix resulting from the use of  $\mathbf{P}_l^{-1}$  rather than  $\mathbf{C}_r$ . Since, in this context,  $\sigma_0^2$  is not known, only an estimate of  $\mathbf{C}_w$  is obtainable, i.e.,

$$\hat{\mathbf{C}}_w = \hat{\sigma}_0^2 \tilde{\mathbf{M}}^{-1}. \quad (12.54)$$

Similar expressions can be obtained for the remaining covariance matrices:

$$\hat{\mathbf{C}}_{\hat{x}} = \hat{\sigma}_0^2 \tilde{\mathbf{N}}^{-1}, \quad (12.55)$$

$$\hat{\mathbf{C}}_{\hat{r}} = \hat{\sigma}_0^2 \mathbf{P}_l^{-1} \mathbf{B}^T \mathbf{L} \mathbf{B} \mathbf{P}_l^{-1}, \quad (12.56)$$

$$\hat{\mathbf{C}}_{\hat{l}} = \mathbf{C}_r - \hat{\mathbf{C}}_{\hat{r}}. \quad (12.57)$$

The most significant impact of using the a posteriori variance factor  $\hat{\sigma}_0^2$  instead of  $\sigma_0^2$  lies within the realm of statistics and is treated fully in Chapter 13.

## CHAPTER 13

### ASSESSMENT OF RESULTS

The objective of this chapter is to show how to statistically assess the quality of the results obtained from overdetermined models using the least-squares method. This is accomplished by demonstrating, in the first section, the interplay of statistics and Hilbert space optimization. The second section deals with the general concepts of statistical testing, applicable equally to the testing of adjusted observations, models, and determined parameters. A description of the individual tests associated with each of the three quantities, i.e., observations, models, and parameters, is given in the third, fourth, and fifth sections.

#### 13.1. Hilbert space and statistics

A systematic method for the study of the behaviour of stochastical quantities involved in experimentation requires the use of *mathematical statistics*. Statistics is used to treat the variables only when a purely deterministic method is unavailable. By bringing together mathematics and the ‘real’ world as sampled through experimentation, conclusions may be reached about the success of the experiment. Nevertheless, a word of caution is in order: statistics should be used with discretion, i.e., in conjunction with common sense, practical experience, and external evidence. Generally, statistics is used only to establish the degree of trustworthiness in the solution. It may, however, also be used to establish whether a determination is compatible with other existing determinations of the same parameters.

The mathematical theory of statistics has its roots in *parametric statistics* [GAUSS, 1809]. In parametric statistics, it is necessary to know or to be able to postulate the form of the probability distribution of the observations  $\mathbf{l}$  needed in the mathematical model before a statistical interpretation of the solution can be given. In particular, the usual statistical techniques to be discussed in this chapter require the observations  $\mathbf{l}$ , or equivalently, the residuals  $\mathbf{r}$ , to be normally distributed. The less this requirement is fulfilled, the less valid are the techniques about to be described. Requirements for normality in statistical techniques should not be confused with the requirements for the least-squares technique discussed in Chapter 12. The latter requires no assumption of normality (if no probabilistic implications are to be drawn).

The term *robust statistics* is used to describe statistical techniques that remain nearly valid even when the statistical quantities investigated are not normally distributed. Though this segment of statistics looks appealing, it will not be applied here because it is not routinely used in geodesy: there is ample experimental evidence that observations in geodesy are most often normally distributed (e.g. BAARD [1967, 1976]), or at least nearly so. Nevertheless, the assumption of normal distribution (normality) should still be tested whenever possible.

*Non-parametric statistics*, on the other hand, does not require any knowledge of the probability distribution of the observations [WONNACOTT AND WOONNACOTT, 1972]. This explains why the term ‘distribution free’ is sometimes used synonymously with the term non-parametric. An extensive bibliography on non-parametric statistics has been compiled by SAVAGE [1953], and a comprehensive text on the subject has been written by SIEGEL [1956]. Here, non-parametric statistics will be referred to only when the parametric statistical approach proves inadequate.

Parametric and non-parametric statistics together are sometimes described as objective, because no subjective a priori information about the unknown parameters in the model is needed before their application. The unknown parameters are treated as if nothing is known about them beforehand, and the solution is arrived at solely on the strength of the observations. When we want the result to reflect a priori subjective knowledge about the parameters, i.e., when a priori subjectively assessed estimates of  $\mathbf{x}$  and  $\mathbf{C}_x$  are used along with the given observations  $\mathbf{l}$  and  $\mathbf{C}_l$ , then *Bayesian statistics*, based on the works of BAYES [1763] and JEFFREYS [1961], are applicable. Some aspects of Bayesian estimation will be treated in §14.4.

To see how statistics fits into the methodology developed thus far, let us begin by interpreting the mathematical model (12.7) and the minimum least-squares distance condition (12.9) from a probabilistic point of view. Clearly, the distance being minimized is the length of the residual vector  $\tilde{\mathbf{r}}$  in the model space  $\mathcal{F}$  which is metricized by the metric tensor  $\mathbf{C}_{\tilde{\mathbf{r}}}^{-1}$ . What does this mean in probabilistic terms? It can be shown that a Hilbert space  $\mathcal{F}$  metricized by  $\mathbf{C}_{\tilde{\mathbf{r}}}^{-1}$  can be regarded as a *probability space* if the inverse of  $\mathbf{C}_{\tilde{\mathbf{r}}}$  can play the role of a *probability measure* in  $\mathcal{F}$  [WILKS, 1962]. In such a probability space, it makes sense to seek probabilistic (statistical) interpretations of the various quantities involved. A complete proof of these statements will not be given here; instead, a heuristic demonstration will be provided of the fact that  $\mathbf{C}_{\tilde{\mathbf{r}}}^{-1}$  can be regarded as a probability measure in  $\mathcal{F}$ .

The matrix  $\mathbf{C}_{\tilde{\mathbf{r}}}^{-1}$  plays the same role, that of a weight matrix, in solving the implicit model as it does in solving the model explicit in  $\mathbf{l}$ . This is clearly seen in the case of a diagonal  $\mathbf{C}_{\tilde{\mathbf{r}}}$  because each residual  $\tilde{r}_i$  in (12.9) contributes to the square of the overall distance  $\tilde{\mathbf{r}}^T \mathbf{C}_{\tilde{\mathbf{r}}}^{-1} \tilde{\mathbf{r}}$  by

$$p_i \tilde{r}_i^2 = \tilde{r}_i^2 / \sigma_i^2. \quad (13.1)$$

Thus, the weight  $p_i$  is inversely proportional to the variance  $\sigma_i^2$ . As known from elementary statistics, this indeed should be the case for *statistical weights*. On the other hand, it is known that experimental statistical weights  $p_i$  of observations belonging to a sample of observations are equal to their experimental probabilities  $p_{ri}$  (see §3.4). Since, in the realm of discrete statistics, there exists this equivalence

between experimental weights and experimental probabilities, it is justifiable to believe that the same relationship exists between the corresponding quantities in the realm of continuous statistics, thereby suggesting that  $1/\sigma_i^2$  in (1) and thus even a diagonal matrix  $\mathbf{C}_r^{-1}$  is a measure of probability. Extension to a non-diagonal  $\mathbf{C}_r^{-1}$  then comes naturally, which concludes the demonstration.

Let us now use the above to enquire into the statistical nature of the least-squares solution of an overdetermined problem. The statistical link between  $\mathcal{F}$  and  $\mathcal{X}$  is secured through the use of the covariance matrix  $\mathbf{C}_r$ . The transformation (cf. eqn. (11.1)),

$$(\tilde{\mathbf{r}}, \mathbf{C}_r) \rightarrow (\hat{\mathbf{x}}, \mathbf{C}_{\hat{\mathbf{x}}}), \quad (13.2)$$

ensures that the statistical information  $\mathbf{C}_r$  in  $\mathcal{F}$ , actually originating in the observation space  $\mathcal{L}$  as  $\mathbf{C}_r = \mathbf{C}_r$ , is transformed into the solution space  $\mathcal{X}$ , by way of the model space  $\mathcal{F}$ . The metric tensor  $\mathbf{C}_{\hat{\mathbf{x}}}^{-1}$  becomes the probability measure in  $\mathcal{X}$ . The other covariance matrices, developed in §12.3, can be statistically interpreted in a similar way.

To illustrate further the relationship between a covariance matrix and probability, let us look at the relationship between the standard deviation and probability. Given the above-mentioned covariance matrices (and thus, among others, the variances for individual elements), what does it mean to say, for instance, that an observable has a value  $\bar{l}$  with a standard deviation of  $\sigma_l$ ? Usually this is understood to mean that the expected value  $\hat{l}$  lies somewhere in the interval  $[\bar{l} - \sigma_l, \bar{l} + \sigma_l]$ , which is often abbreviated as  $\bar{l} \pm \sigma_l$ . How trustworthy is this statement? Because measurements cannot be made without errors, complete certainty can never be achieved. To compensate for the everpresent uncertainty, an attempt is made to assign a value to the probability that our statement  $\hat{l} \in (\bar{l} \pm \sigma_l)$  is correct. But how is such a probability value obtained, and how is it utilized? If a probability can be assigned to such a statement involving the observations, is it then possible to do the same for the unknown parameters?

The residuals  $\mathbf{r}$ , i.e., the stochastical part of the observation, are used to provide answers to these questions in the same way that  $\mathbf{r}$  were used to evaluate  $\mathbf{C}_r$ . Without any loss of generality, it is assumed for this purpose that the data series,  $l(\tau_i)$ ,  $i = 1, \dots, N$ , is composed of a constant ( $l(\tau_i) = \text{const.}$ ) and the residual series ( $r(\tau_i)$ ). Further, it is assumed that the residuals are statistically independent, i.e.,  $r(\tau) = v(\tau)$ . Then a histogram or polygon of the observations,  $l(\tau_i)$ ,  $i = 1, \dots, N$ , can be plotted (see FIG. 1) showing the distribution of the experimental probabilities.

Under certain conditions, and for some observables, it is possible to make a large number of repeated measurements resulting in a construction of a smoother, more meaningful histogram. Such numerous measurements though are usually made only during instrument calibration; it is not economically feasible, for routine operations, to collect a wealth of data solely for the purpose of constructing a smoother histogram. Still, histograms are merely a geometrical representation of numerical (discrete) functions, which, in turn, are inconvenient to work with analytically. It is more convenient to work with a compact idealization of the 'discrete world of experimentation'. For this reason, a smooth curve is postulated to approximate the

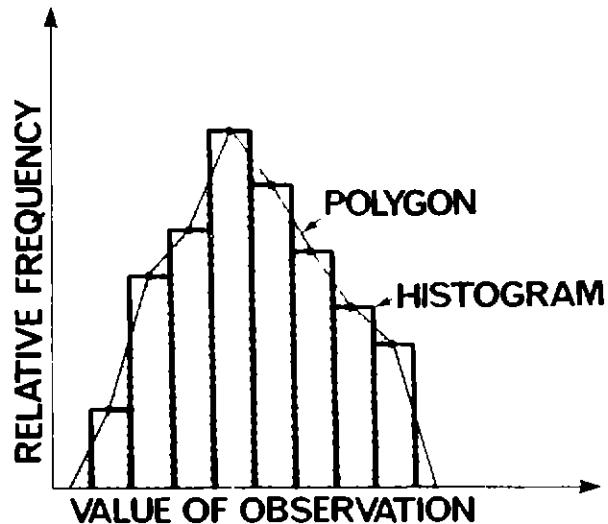


FIG. 13.1. Polygon and histogram.

histogram (polygon). The assumption justifying this approximation is that there exists an infinite population with a smooth histogram (polygon) of which our  $N$  observations are only a sample. This is the *basic postulate of mathematical statistics*.

Because the histogram can be regarded as depicting the distribution of experimental probabilities, it is natural to apply this meaning also to the smooth curve and view the curve as a postulated probability density function (see §3.4). Using the postulated probability density function, the answer to the question posed earlier ( $\text{pr}(\bar{l} - \sigma_l < \hat{l} < \bar{l} + \sigma_l) = ?$ ) is (cf. FIG. 2)

$$\text{pr}(\bar{l} - \sigma_l < \hat{l} < \bar{l} + \sigma_l) = \int_{\bar{l} - \sigma_l}^{\bar{l} + \sigma_l} \phi_l(\xi) d\xi, \quad (13.3)$$

where  $\sigma_l^2$  is now interpreted as the variance of the postulated probability density function (see §3.4). The ultimate goal, to be able to write the probability statement (3) for the unknown parameters as well, cannot be tackled yet, however, and will

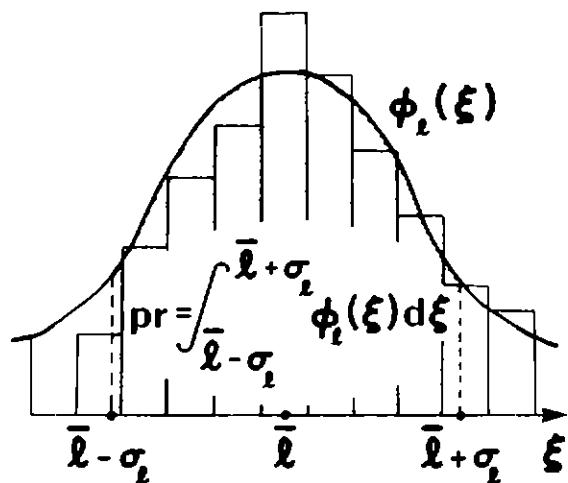


FIG. 13.2. Histogram and probability density function.

have to wait till §13.5. Before it can be tackled, the statistical ideas presented thus far have to be further formalized and extended to the multidimensional situation, because the components of  $x$  are jointly related to all the observations, i.e., to the whole vector  $\boldsymbol{l}$ .

In hypothesizing a probability density function that underlies the histogram of observations  $\boldsymbol{l}$ , both its mathematical form  $\phi_l$  and the values of its parameters must be postulated. The knowledge of the postulated form and the parameters, called in this context *population parameters*  $\theta_i$ ,  $i = 1, 2, \dots$ , must come from sources external to the reasoning followed here. If there is no basis for the postulation of the population parameters—the form of  $\phi_l$  must, however, always be postulated—then their values can be estimated from the sample.

To show how this is done, let us restrict ourselves to dealing with only two-parametric probability density functions and assume further that  $\theta_1 = \mu$  and  $\theta_2 = \sigma^2$  (see §3.4). In practice, the fewer parameters the better and the reduction of the two parameters to the mean and variance can always be achieved. Then the assertion of ‘known  $\theta_1$ ’ is equivalent to postulating the expected value  $\bar{l}$  of  $\boldsymbol{l}$  to be equal to  $\mu_l = \theta_1^{(l)}$ , and similarly ‘known  $\theta_2$ ’ means  $\sigma_l^2 = \theta_2^{(l)}$ . When there is no basis for the postulation of  $\theta_i^{(l)}$ , i.e., in the case of ‘unknown  $\theta_i$ ’, the postulated probability density function is of the form  $\phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)})$ , where  $\hat{\theta}_1^{(l)} = \bar{l}$  and  $\hat{\theta}_2^{(l)} = s_l^2$  (see §3.4). Of course, combinations of both cases are possible, e.g.,  $\phi_l(\xi; \theta_1^{(l)}, \hat{\theta}_2^{(l)})$ .

The use of either known values or the estimated values of the parameters does not automatically ensure that the postulated probability density function adequately describes reality, i.e., the histogram. The question then arises, is the probability value  $pr = \int_a^b \phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)}) d\xi$  really to be trusted? Some of the things that can go wrong are illustrated in FIGS. 3 and 4. It is obvious that the depicted postulated probability density functions are inconsistent with the histogram, and thus the use of  $\phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)})$  in the calculations of probability can lead to incorrect statements about the statistical behaviour of the observations. When  $\phi_l(\xi; \theta_1^{(l)} = \hat{\theta}_1^{(l)}, \theta_2^{(l)})$  is

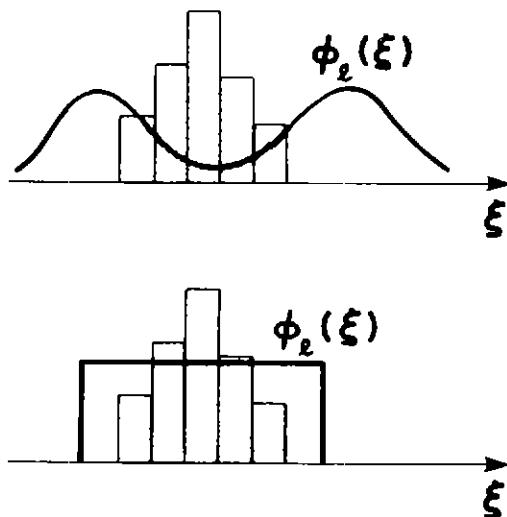
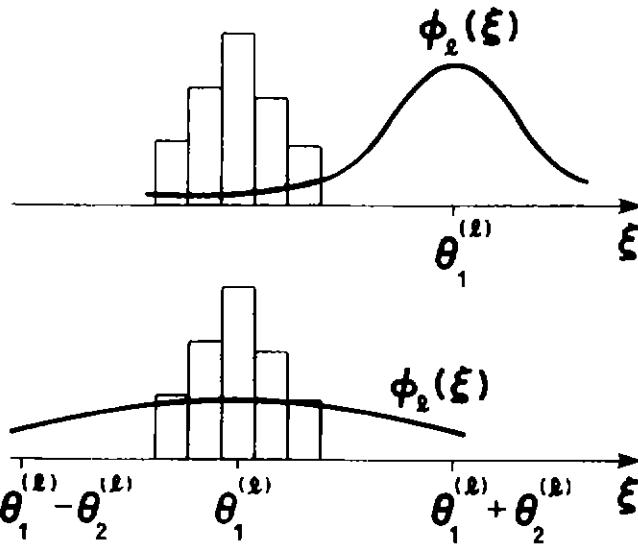


FIG. 13.3. Examples of incorrect postulation of the shape of  $\phi_l(\xi)$ .

FIG. 13.4. Examples of incorrect postulations of  $\theta_1^{(1)}, \theta_2^{(1)}$ .

postulated—‘correctly’ as far as  $\theta_1^{(1)}$  is concerned—it is referred to as *unbiased with respect to  $\theta_1$* ; this situation is usually described in the literature as ‘ $\hat{\theta}_1$  is an unbiased estimate of  $\theta_1$  because  $E(\theta_1) = \hat{\theta}_1$ ’ (e.g., HAMILTON [1967]). The equivalence of these two approaches is seen immediately when  $E(\theta_1) = \int_{-\infty}^{\infty} \xi \phi(\xi; \hat{\theta}_1, \theta_2) d\xi = \hat{\theta}_1$  is evaluated (cf. §3.4). Similarly, a probability density function is known as *unbiased with respect to  $\theta_2$* , if one postulates  $\phi_1(\xi; \theta_1^{(1)}, \hat{\theta}_2^{(1)})$ . If the postulates are not unbiased then they are biased. Note, for instance, that if  $\phi_l(\xi; \hat{l}, \sigma_l^2)$  is postulated to be the probability density function of  $l$ , then this postulate is unbiased, since  $\hat{l} = \bar{l}$ . Statistical testing, in its broadest sense, is a means of helping to discover whether anything has gone wrong with the basic postulate. A variety of tests can be performed in the univariate or multivariate situations. These depend upon what specifically is being tested: the shape of the probability density function, the postulated population parameters, or both.

In the multivariate situation, the estimates ( $\hat{l}$ ,  $\hat{r}$ , and  $\hat{x}$ ) for the expected values of observables, residuals, and unknown parameters, and the covariance matrices ( $C_{\hat{l}}$ ,  $C_{\hat{r}}$ , and  $C_{\hat{x}}$ ) for the accuracy of these three (estimated) quantities have been derived in Chapter 12. Taken in pairs, i.e.,  $(\hat{l}, C_{\hat{l}})$ ,  $(\hat{r}, C_{\hat{r}})$ , and  $(\hat{x}, C_{\hat{x}})$ , clearly, each pair is an estimate of the two *multidimensional population parameters* ( $\theta_1, \theta_2$ ) to be used in their respective postulated multidimensional probability density functions:  $\phi_l(\xi; \hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$ ,  $\phi_r(\xi; \hat{\theta}_1^{(r)}, \hat{\theta}_2^{(r)})$ ,  $\phi_x(\xi; \hat{\theta}_1^{(x)}, \hat{\theta}_2^{(x)})$ .

It is useful to finish this section by stating the following *properties of the least-squares solution* to overdetermined models, as given in HAMILTON [1967] and GRAYBILL [1976]:

- (a) The least-squares estimate  $\hat{x}$  (12.11) of  $\hat{x}$  is unique, if  $N$  (12.23) is non-singular.
- (b)  $\hat{x}$  is an unbiased estimate of the expected value  $\hat{x}$  of  $x$ , if  $E(r) = 0$  (cf. (10.42)).
- (c)  $\hat{x}$  is a minimum variance estimate of the expected value of  $x$ , in the sense that the trace of  $C_{\hat{x}}$  (12.36) is a minimum.

- (d)  $\hat{x}$  is a maximum likelihood estimate of the expected value of  $x$ , if the residuals  $r$  have a normal probability density distribution.
- (e)  $\hat{r}$  (12.37) is an unbiased estimate of the expected value  $\bar{r}$  of  $r$ .
- (f)  $\hat{\sigma}_0^2$  (12.51) is an unbiased estimate of the expected value of  $\sigma_0^2$ , thus making  $\hat{C}_{\hat{x}}$  (12.55)  $\hat{C}_{\hat{r}}$  (12.56), and  $\hat{C}_{\hat{l}}$  (12.57) respectively unbiased estimates of expected  $C_{\hat{x}}$ ,  $C_{\hat{r}}$ , and  $C_{\hat{l}}$ . This holds true only under the condition that the weight matrix in eqn. (11.37) has been selected to be equal to the inverse of the covariance matrix of the observations (i.e.,  $P = C_l^{-1}$ ). This is the rationale for the selection made earlier in §11.4 and unexplained until now.

### 13.2. Statistical testing

The role of statistical testing is to determine whether or not:

- (a) the postulated probability density function for the experiment (sample) is likely to have been correctly postulated;
- (b) the estimated value of a population parameter is to be trusted; and
- (c) the estimated value of a population parameter is consistent with the known (a priori) value of the parameter, if it is available.

Before discussing the testing further, several definitions must be introduced. A *statistic* is a special random variable that is a function of one or more random variables (cf. §3.4), and it does not depend on any unknown population parameter. Here, the statistics  $y$  are functions of the population parameters of the probability density function belonging to the observables or parameters. The probability density function  $\phi$  of  $y$  is derived from the probability density function of, say,  $l$  as follows [HAMILTON, 1967]:

$$\phi_y(\xi; \theta_1^{(y)}, \theta_2^{(y)}) = \det J \phi_l(\xi; \theta_1^{(l)}, \theta_2^{(l)}), \quad (13.4)$$

where  $J$  is the Jacobian matrix of transformation (see §3.1) from the probability space  $\mathcal{L}$  (containing  $l$ ) into the probability space  $\mathcal{Y}$  (containing  $y$ ), and  $\phi_l$  is the postulated probability density function of  $l$ . An identical relationship exists for the multivariate case (cf. §13.4).

A *statistical hypothesis* is a quantitative statement about the postulated (hypothesized) probability density function and its parameters. If, in the hypothesis, the postulated density function is completely specified, then one speaks about a *simple statistical hypothesis* [HOGG AND CRAIG, 1970]. If it is not specified, then reference is made to a *composite statistical hypothesis*. If the population parameters  $\theta_1, \theta_2$  are postulated (hypothesized) to have some particular values  $\theta_1, \theta_2$ , then this is called the *null hypothesis*  $H_0$ . A null hypothesis can be restrictive, i.e., specify only one population parameter, or general, i.e., specify particular values for all population parameters. For every null hypothesis, there is an infinite number of *alternative hypotheses*, each of which states that the population parameters have some other particular values. Logically, an alternative hypothesis may also be simple or composite; in our studies, however, only simple hypotheses will be used. Any hypothesis

$H$  has a statistic  $y$  associated with it. The probability density function of such a statistic will be denoted by  $\phi_{y/H}$ , or by  $\phi_y(\xi; \theta_1^{(y)}, \theta_2^{(y)})$ , or simply just by  $\phi_y$ .

Any statistical hypothesis can be tested: a *test of a statistical hypothesis*  $H_0$  is an algorithm that leads to a statistical decision concerning the validity of  $H_0$ . A complete agreement between the discrete and compact approaches, and, consequently, a definite decision concerning  $H_0$ , can be reached only on the basis of an infinite sample which, of course, is never available. A decision based on a finite sample can be trusted only to a certain degree. This means that such a decision has only a limited confidence attached to it.

There are two possible outcomes of the test: 'accept  $H_0$ ' or 'reject  $H_0$ '. Similarly, there are two possible outcomes of the same test for an alternative hypothesis  $H_1$ . Now, since none of the hypotheses may be true, the test should at least show which hypothesis is better. The problem is that if, say,  $H_0$  is not true, there is no guarantee that  $H_1$  is true. The situation is summarized in TABLE 1. The probability  $\alpha$  of rejecting  $H_0$  when, in fact,  $H_0$  is true (Type I error) is called the *significance level*. The value of  $\alpha$  that must be selected before the test is carried out has to lie between 0 and 1, and should be as small as possible. However, for finite samples, one discovers that no  $H_0$  is acceptable if  $\alpha=0$ , i.e., if there is no risk involved. In geodesy, values between 0.01 and 0.05 are usually selected. The complementary probability  $1-\alpha$  is called the *confidence level*, and it is the measure of confidence to be had in the decision. It is obvious that, once  $\alpha$  is selected,  $1-\alpha$  is obtained automatically. Shown in FIG. 5 is the relationship between the two probabilities and the role of the probability density function of the statistic  $y$  used to test  $H_0$ . Note that  $\xi_{y,1-\alpha} = \xi_{\phi_y,1-\alpha}$  is the value of the abscissa, which is fixed through the selected value of  $\alpha$ . If  $H_0$  is a simple hypothesis, then the following can be written for the two areas under the curve  $\phi_{y/H_0} = \phi_y(\xi; \theta_1^{(y)}, \theta_2^{(y)})$ :

$$\text{pr}(y > \xi_{y,1-\alpha}) = \int_{\xi_{y,1-\alpha}}^{\infty} \phi_y(\xi; \theta_1^{(y)}, \theta_2^{(y)}) d\xi = \alpha, \quad (13.5)$$

and

$$\text{pr}(y < \xi_{y,1-\alpha}) = \int_{-\infty}^{\xi_{y,1-\alpha}} \phi_y(\xi; \theta_1^{(y)}, \theta_2^{(y)}) d\xi = 1 - \alpha.$$

Let us now examine the situation when  $H_1$  is true or, in other words, when  $H_0$  is false. Again there are two possible outcomes: 'accept  $H_0$ ' or 'reject  $H_0$ '. The

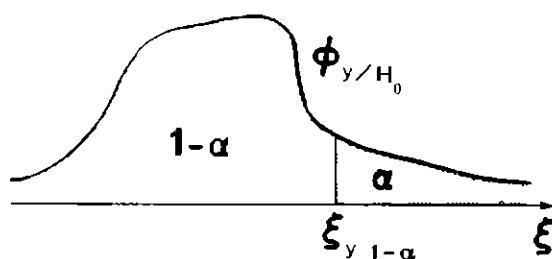


FIG. 13.5. Significance level ( $\alpha$ ) and confidence level ( $1-\alpha$ ).

TABLE 13.1  
Testing of a null hypothesis  $H_0$  against an alternative hypothesis  $H_1$

Decision Situation \n	Test tells us to accept $H_0$	Test tells us to reject $H_0$
$H_0$ true	Correct decision: $\text{pr} = 1 - \alpha$ (confidence level)	Type I error: $\text{pr} = \alpha$ (significance level)
$H_0$ false ( $H_1$ true)	Type II error: $\text{pr} = \beta$	Correct decision: $\text{pr} = 1 - \beta$ (power)

probability  $\beta$  of accepting  $H_0$  when it is false (Type II error) is related to the *power of the test*  $1 - \beta$ , in a complementary fashion. For the same reasons as above,  $\beta$  should be as small as possible. A value for  $\beta$  can only be computed if an alternative, simple hypothesis  $H_1$  is put forward (with a completely specified probability density function of its statistic  $y_1$ , i.e.,  $\phi_{y_1/H_1} = \phi_{y_1}(\xi; \theta_1^{(y_1)}, \theta_2^{(y_1)})$ ). Usually, the probability density function of  $y_1$  has the same form as that of  $y$ ; they would typically differ only by having different values for  $\theta_1, \theta_2$ . To be useful, the alternative hypothesis  $H_1$  must be a reasonable one so that both alternatives are plausible. This means that  $H_1$  should be appreciably, but not unreasonably, different from  $H_0$ . FIG. 6 shows a reasonable  $H_1$  that somewhat overlaps with  $H_0$ , as viewed through their probability density functions. The equation for the area  $\beta$  can be written as

$$\text{pr}(y_1 < \xi_{y_1, 1-\alpha}) = \int_{-\infty}^{\xi_{y_1, 1-\alpha}} \phi_{y_1}(\xi; \theta_1^{(y_1)}, \theta_2^{(y_1)}) d\xi = \beta. \quad (13.6)$$

Returning to the relationship between  $\alpha$  and  $\beta$ , note that a decrease in  $\alpha$  leads to an increase in  $\beta$  and vice-versa; thus a compromise must result. Evidently, both  $H_0$  and  $H_1$  must be simple hypotheses to enable us to do the above. The *most powerful test* is the one employing the particular alternative hypothesis  $H_1$  that yields the smallest Type II error ( $\beta$ ) for the same significance level  $\alpha$ . In FIG. 7, it is clear that, of the two alternative hypotheses  $H_1, H_2$  shown, the test involving  $H_1$  is better than the one involving  $H_2$ , because for a chosen  $\alpha$  (i.e.,  $\xi_{y_1, 1-\alpha}$ ),  $\beta_1 < \beta_2$ .

A mathematically equivalent alternative [WONNACOTT AND WONNACOTT, 1972] to statistical testing is the concept of *confidence regions*. It is used much more often in geodesy than statistical testing because, in geodesy, it is seldom possible to present

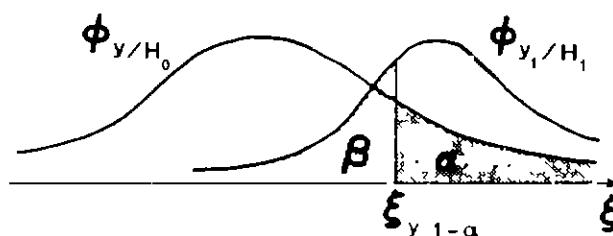


FIG. 13.6. Interplay of Type I and Type II errors.

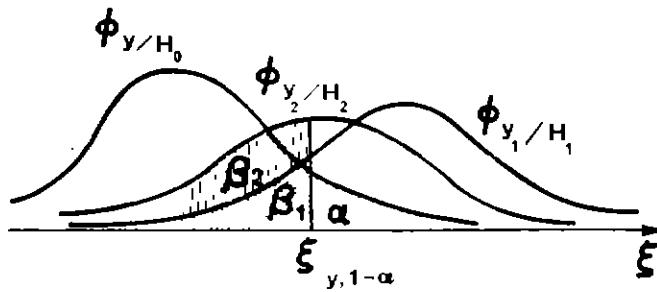


FIG. 13.7. The most powerful test.

two clear-cut alternative hypotheses with the actual situation being always somewhere between the two. To explain this concept, let us begin with formulating the simple null hypothesis  $H_0$ : ' $\theta_1^{(I)} = \theta_1$ , and  $\theta_2^{(I)} = \theta_2$ '. To test this hypothesis, let us design a statistic,

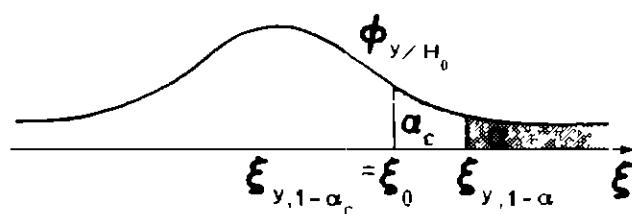
$$y = y(\phi_I; \theta_1^{(I)}, \theta_2^{(I)}), \quad (13.7)$$

with a probability density function  $\phi_y(\xi; \theta_1^{(y)}, \theta_2^{(y)})$ . Evidently, for any particular values  $\theta_1, \theta_2$  there is a corresponding particular value of  $y$ ; let us denote this value  $y(\phi_I; \theta_1, \theta_2)$  by  $\xi_0$ . The probability  $1 - \alpha_c$  that the above  $H_0$  is correct can now be evaluated. As seen above,

$$\text{pr}(y > \xi_0) = \int_{\xi_0}^{\infty} \phi_y(\xi; \theta_1^{(y)}, \theta_2^{(y)}) d\xi = \alpha_c. \quad (13.8)$$

The actual probability value  $\alpha_c$ , which is only a function of  $\theta_1$  and  $\theta_2$  through the selected statistic  $y$ , is called the *critical significance level*. The complement  $1 - \alpha_c$  is known as the *critical confidence level*. Evidently, if  $\alpha_c > \alpha$ , then  $H_0$  is acceptable at the  $\alpha$  significance level (see FIG. 8). If  $\alpha = \alpha_c$ , then the acceptability on the  $\alpha$  level of significance is questionable—hence the name critical. Of course, if  $\alpha_c < \alpha$ , then  $H_0$  is not acceptable on the  $\alpha$  significance level.

So far it has been tacitly assumed that the statistic  $y$  was selected so that only its upper (right-hand side) boundary mattered for the assessment of the validity of  $H_0$ . Thus, *one-sided probability values*  $\alpha$  have been dealt with. The more general case, where both the lower and upper boundaries matter, is also encountered in practice. Clearly, if the hypothesis calls for the value of  $\theta_1^{(y)}$  to equal  $\theta_1$ , the values smaller than  $\theta_1$  are, from the point of view of the hypothesis, as incorrect as those

FIG. 13.8. Critical significance level  $\alpha_c$  and significance level  $\alpha$  (one-tail).

larger than  $\theta_1$ . This situation is depicted in FIG. 9; the probability density function of the  $y$  being used may be either symmetrical (a) or non-symmetrical (b) depending on the formulation of the statistic. Accordingly, one speaks about *two-sided* (tailed) *probability values*, and the corresponding probability statement is

$$\text{pr}(\xi_{y,\alpha/2} < y < \xi_{y,1-\alpha/2}) = 1 - \alpha, \quad (13.9)$$

where, normally, half of the required probability value  $\alpha$  is taken to be contained in the right-hand side tail and the other half in the other tail. The interval  $[\xi_{y,\alpha/2}, \xi_{y,1-\alpha/2}]$  is naturally called the confidence interval (of confidence  $1 - \alpha$ ) or, briefly, the  $(1 - \alpha)$  *confidence interval* (see FIG. 9(c)). The critical value  $\alpha_c$  is treated in a similar way; namely, for a *two-tailed test*,  $y_{\alpha_c/2}$  is equated to  $\xi_0$  and  $\frac{1}{2}\alpha_c$  is obtained. Then  $\alpha_c$  is compared with  $\alpha$ , as explained earlier for the *one-tail test*. The probability statement given by (9) establishes the corresponding  $(1 - \alpha)$  confidence interval

$$\boxed{\xi_{y,\alpha/2} < y < \xi_{y,1-\alpha/2}} \quad (13.10)$$

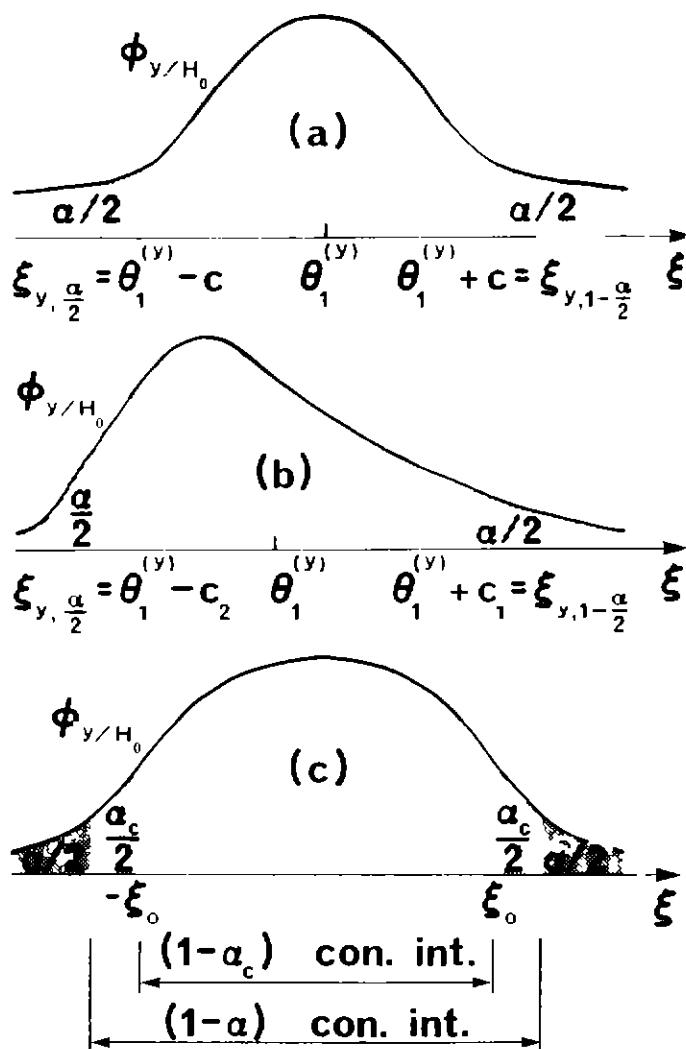


FIG. 13.9. Two-tailed confidence interval.

If the  $y_0$  fails to fall inside this interval,  $H_0$  is rejected at the prescribed confidence level  $1 - \alpha$ . The advantage of the confidence interval approach, over hypothesis testing, is that one can see by how much  $\xi_0$ , as a function of the hypothesized  $\theta_i$ , falls inside or outside the interval.

When the hypothesis tested is  $H_0: \theta_i = \hat{\theta}_i$ , for  $i = 1$  or  $2$  or both of them, the confidence interval can usually be reformulated to hold for the  $\theta_i$  rather than  $y$ . This approach is widely used in geodesy, as will be shown further on.

### 13.3. Assessment of observations of one observable

The rationale behind assessing the observations is quality control. Good observations produce good results, i.e., trustworthy values for the parameters, while bad observations produce bad results. Of course, even good observations cannot produce good results if the mathematical model is not formulated correctly. As already discussed (steps (d) and (f) in §10.1), observations are assessed on two occasions—after they have been procured, and after they have been introduced into the model. Thus, there are two different, though related, sets of tests. In this section, the tests of the observations are treated on their own; in the next section the tests assessing the observations are treated in the light of their fit to the model.

Let us begin by examining the residual  $r$  in more detail. It is made up of two components—the  $v$  and the  $s$  residuals. Because the univariate probability density function contains only one variance  $\sigma^2$  and no covariances,  $s$ , statistically dependent by definition, cannot be incorporated into the univariate scheme. Multivariate techniques are needed, as will be seen in §13.4. In this section, it is thus assumed that  $l$  can be decomposed as follows:  $l = t + r$ , where  $t = \mu_i = \text{const.}$ , and  $r = v$ . Even under these simplifying assumptions, the observations (data) series can be tested from two different points of view, both using *univariate tests*. One point of view treats the series  $l(\tau_i)$  as a whole and examines the correctness of the probability density function postulated for it. The second examines each observed value separately and asks if the observation is compatible with the rest of the values. In both families of tests, it is highly recommended to combine the (objective) statistical assessment with the subjective assessment using the recorded history of the series; i.e., readings, time, conditions, instrumentation, personnel, etc.

The tests for the entire observational series are summarized in TABLE 2. Six situations, often found in practice, are represented. They stem from whether the population mean  $\mu$  and population variance  $\sigma^2$  are ‘known’ or ‘unknown’; when ‘unknown’ they are estimated by the sample mean  $\bar{l}$  and sample variance  $s^2$ . The following example can be taken to illustrate the differences:

- (a) ‘ $\mu$  known’ corresponds to measuring a known distance;
- (b) ‘ $\sigma^2$  known’ corresponds to measuring with an instrument of known accuracy;
- (c) ‘ $\mu$  unknown’ corresponds to measuring a line of unknown length; and
- (d) ‘ $\sigma^2$  unknown’ corresponds to measuring a line with an instrument of unknown accuracy.

TABLE I 3.2  
Assessment of an observation series as a unit

Name	Situation		$H_0$ (null hypothesis)		Statistic	Probability	
	$\theta_1$	$\theta_2$	$y$	function of $y^a$	density function of $y^a$	$1-\alpha$ confidence interval for the quantity being tested	Remarks
$\chi^2$ goodness of fit test	1	$\mu$ known	$\sigma^2$ known	histogram compatible with $n(\xi; \mu, \sigma^2)$	$\sum_{i=1}^n \frac{(a_i - e_i)^2}{e_i}$	$\chi^2(\xi; n-1)$	$0 < y < \xi_{\chi_{n-1}^2, 1-\alpha}$
	2	unknown	unknown	histogram compatible with $n(\xi; \bar{t}, s^2)$	$\sum_{i=1}^n \frac{(a_i - e_i)^2}{e_i}$	$\chi^2(\xi; n-3)$	$0 < y < \xi_{\chi_{n-3}^2, 1-\alpha}$
		$\mu$	$\sigma^2$				
		$\bar{t}$	$s^2$				
		used	used				
Test on the variance	3	$\mu$ known	$\sigma^2$ known	sample has the probability density function $n(\xi; \mu, \sigma^2)$	$\sum_1^N \left( \frac{l_i - \mu}{\sigma} \right)^2 = \frac{Ns^2}{\sigma^2}$	$\chi^2(\xi; N)$	$\frac{Ns^2}{\xi_{\chi_{N-1-\alpha/2}^2}} < \sigma^2 < \frac{Ns^2}{\xi_{\chi_{N-\alpha/2}^2}}$
	4	unknown	$\bar{t}$ (tested)	sample has the probability density function $n(\xi; \bar{t}, \sigma^2)$	$\sum_1^N \left( \frac{l_i - \bar{t}}{\sigma} \right)^2 = \frac{(N-1)s^2}{\sigma^2}$	$\chi^2(\xi; N-1)$	$\frac{(N-1)s^2}{\xi_{\chi_{N-1-\alpha/2}^2}} < \sigma^2 < \frac{(N-1)s^2}{\xi_{\chi_{N-1-\alpha/2}^2}}$
		$\mu$					
		$\bar{t}$	(tested)				
		used					
Test on the mean	5	$\mu$ known	$\sigma^2$ known	sample has the probability density function $n(\xi; \mu, \sigma^2)$	$\frac{\bar{l} - \mu}{\sigma / \sqrt{N}}$	$n(\xi; 0, 1)$	$\bar{l} - \frac{\sigma}{\sqrt{N}} \xi_{n(0, 1), 1-\alpha/2} < \mu < \bar{l} + \frac{\sigma}{\sqrt{N}} \xi_{n(0, 1), 1-\alpha/2}$
	6	known	$\bar{t}$ (tested)	sample has the probability density function $n(\xi; \bar{t}, s^2)$	$\frac{\bar{l} - \mu}{s / \sqrt{N}}$	$t(\xi; N-1)$	$\bar{l} - \frac{s}{\sqrt{N}} \xi_{t_{N-1}, 1-\alpha/2} < \mu < \bar{l} + \frac{s}{\sqrt{N}} \xi_{t_{N-1}, 1-\alpha/2}$
		$\mu$					
		$\bar{t}$	(tested)				
		used					

<sup>a</sup> HOGG AND CRAIG [1970]:  $\chi^2(\xi; v)$ -chi-squared density with  $v$  degrees of freedom.

$n(\xi, 0, 1)$ -standard normal density with a mean of 0 and a variance of 1.

$t(\xi; v)$ -Student's  $t$  density with  $v$  degrees of freedom.

$n$ =number of classes.  
 $a_i$ =actual count.  
 $e_i$ =theoretical count.

Loss of two degrees of freedom since sample mean and variance are used.

$N$ =sample size; to test if  $s^2 = \sigma^2$ .

To test if  $\bar{l} = \mu$ .

$\bar{l}$  and  $v$  must be computed from independent samples.

The main objective of the  $\chi^2$  *goodness of fit test* (tests numbered 1 and 2) is to test if the histogram is compatible with a postulated probability density function, usually the normal. This is a crucial test because all the other tests are based on the assumption of normality. But the test itself is equally valid for other probability density functions.

The  $\chi^2$  *test on the variance* (tests numbered 3 and 4) determines whether the hypothesized population variance ( $\sigma^2$ ) is indeed compatible with the value  $s^2$  estimated from the sample. Sometimes,  $\sigma^2$  may be viewed as the *design variance*, i.e., the variance of the observations required to achieve a certain accuracy in the parameters (step (c) in §10.1), and  $s^2$  as the variance of the sample observations. If the test shows that  $s^2$  and  $\sigma^2$  are incompatible (test fails), then it can be said that there is evidence suggesting that the observations have not been collected according to the design, and, thus, a reassessment of the measurement process is indicated. Even if the test passes, however,  $\sigma^2$  is used to characterize the accuracy of the data in subsequent work. In other words, this test may serve as a means of calibrating the accuracy of the measuring process or instrument.

The *normal test* and *Student's t test on the mean* (tests numbered 5 and 6) are designed to examine the mean of a data series. The two tests check for the compatibility of  $\mu$  and  $\bar{l}$ , i.e., for the presence of possible bias in the observed sample.

The tests listed in TABLE 2 may fail because of

- (a) a lack of normality of the histogram;
- (b) incompatibility of  $s^2$  with  $\sigma^2$ , or  $\bar{l}$  with  $\mu$ , or both of them;
- (c) the presence of as yet unaccounted for systematic errors in  $l$ ; and
- (d) the statistical dependence of individual  $l$ 's, i.e., the presence of the statistically dependent residual  $s$ .

There are tests, other than those summarized in TABLE 2, that can be used to assess the observation series  $l(\tau_i)$ ,  $i = 1, \dots, N$ . The interested reader is referred to the literature (e.g., CROW ET AL. [1960], WONNACOTT AND WONNACOTT [1972]) for tests that can be used to verify the compatibility of two or more observational series of the same observable  $l_j$ .

Let us focus now on the individual observation  $l(\tau_i)$  of a series in an attempt to identify and reject *outliers*, i.e., observations which are considered statistically incompatible with the rest of the series. This incompatibility is usually caused by a blunder made in the measurement or by some sort of instantaneous disturbance affecting the performance of the measuring system. Numerous authors have discussed the rejection of outliers (e.g., CHAUVENET [1871], WILLKE [1965], HAMILTON [1967], POPE [1976]), and a synthesis of their work, in terms of the most straightforward tests that exist for the assessment of individual observations, is presented here. All the tests have the same underlying assumption of normality and are summarized in TABLE 3. The interested reader is referred to the literature, which is rich with other, more specialized, tests (e.g., QUESENBERRY AND DAVID [1961], DIXON [1962]).

Again, four distinctly different situations may occur that parallel the last four given in TABLE 2. Practically speaking, these situations may have the same meaning as before. Assumed in TABLE 3 are the definitions of the expected residual and the

TABLE 13.3  
Tests for outliers (univariate case)

Name	Situation		$H_0$ (null hypothesis)	Statistic	Probability density function of $y^*$	$1 - \alpha$ confidence interval for the quantity being tested <sup>b</sup>	Remarks
	$\theta_1$	$\theta_2$					
Normal test of a single observation	1	$\sigma^2$ known	$\bar{l}$ has the probability density function	$\frac{\bar{l} - \mu}{\sigma}$	$n(\xi; 0, 1)$	$\mu - \sigma \xi_{n(0, 1), 1-\alpha/2} < \bar{l}_i$ $< \mu + \sigma \xi_{n(0, 1), 1-\alpha/2}$	$\sigma^2$ known thus the normal density.
Student's $t$ test of a single observation	2	$s^2$ used	$\bar{l}$ has the probability density function	$\frac{\bar{l} - \mu}{s}$	$r(\xi; N-1)$	$\mu - s \xi_{r_{N-1}, 1-\alpha/2} < \bar{l}_i$ $< \mu + s \xi_{r_{N-1}, 1-\alpha/2}$	The tested $\bar{l}_i$ must not have been used in computing $s$ .
Normal test of a single observation	3	$\mu$ unknown	$\bar{l}$ has the probability density function	$\left(\frac{N-1}{N}\right)^{1/2} \sigma$	$n(\xi; 0, 1)$	$\bar{l} - \left(\frac{N-1}{N}\right)^{1/2} \sigma \xi_{n(0, 1), 1-\alpha/2} < \bar{l}_i$ $< \bar{l} + \left(\frac{N-1}{N}\right)^{1/2} \sigma \xi_{n(0, 1), 1-\alpha/2}$	$\sigma^2$ known thus the normal density.
$\tau$ test of a single observation	4	$\bar{l}$ used	$\bar{l}$ has the probability density function	$\left(\frac{N-1}{N}\right)^{1/2} s$	$r(\xi; N-1)$	$\bar{l} - \left(\frac{N-1}{N}\right)^{1/2} s \xi_{r_{N-1}, 1-\alpha/2} < \bar{l}_i$ $< \bar{l} + \left(\frac{N-1}{N}\right)^{1/2} s \xi_{r_{N-1}, 1-\alpha/2}$	$\bar{l}$ and $s^2$ computed from the same sample thus the $\tau$ density.

<sup>a</sup>POP: [1976].  $n(\xi; 0, 1)$ —standard normal density with a mean of 0 and a variance of 1.

$t(\xi; N-1)$ —Student's  $t$  density with  $N-1$  degrees of freedom.

$\tau(\xi; N-1)$ —tau density with  $N-1$  degrees of freedom.

<sup>b</sup>For in-context testing of the series: Replace  $\alpha$  with  $\alpha/N$ , where  $N$  is the number of elements in the series.

estimated residual: the expected residual,  $\hat{v}_i = l_i - \mu$ , is found in the first two statistics, while the estimated residual,  $\hat{v}_i = l_i - \bar{l}$ , is found in the last two statistics. It is instructive to point out that the statistics are simply *standardized residuals* (the role of standardization will be explained in detail in §13.4); in each case, the denominator is merely the standard deviation of the residuals. In the first two cases this fact is obvious. The proof that it is true, even for the third and fourth statistics, lies with the covariance matrix of the estimated residuals given by (12.38). Assuming, for simplicity, the model explicit in  $l$ , i.e.,  $B = -I$ , one gets

$$C_v = C_v - AC_{\hat{x}}A^T. \quad (13.11)$$

Taking now  $C_v = \sigma^2 I$ , the design matrix  $A = [1, 1, \dots, 1]$ , and the covariance matrix  $C_{\hat{x}}$  of the estimated unknown (i.e., of the mean  $\bar{l}$  which causes  $C_{\hat{x}}$  to degenerate to  $C_{\bar{l}} = \sigma^2/N$ ), one gets after multiplication

$$C_v = \sigma^2 \begin{bmatrix} (N-1)/N & -1/N & \cdots & -1/N \\ -1/N & (N-1)/N & \cdots & -1/N \\ \vdots & & & \\ -1/N & -1/N & \cdots & (N-1)/N \end{bmatrix}. \quad (13.12)$$

Clearly, the variance of the  $i$ th estimated residual is  $(N-1)/N\sigma^2$  as claimed above. When  $\sigma^2$  is unknown, as for the fourth statistic in TABLE 3, the variance of the  $i$ th estimated residual is  $(N-1)/Ns^2$ . Equation (12) also shows that, unlike the expected residuals, the estimated residuals are statistically dependent. For the time being, the covariances  $\sigma^2/N$  and  $s^2/N$  cannot be accommodated because only one residual is being examined at a time.

In the four tests given in TABLE 3, each  $l_i$  has actually been taken out of context, and the existence of the other members of the series has deliberately been disregarded. This is why these tests are called *out-of-context tests*. The individual  $l_i$ , or  $r_i$ , can also be examined in the context of their being members of the series, and the *in-context tests* will be now introduced. In doing this, it is expedient to work with standardized residuals, i.e., the statistics  $y$  in TABLE 3, rather than the residuals as such. Since the aim is to seek cut-off points below and above which the (standardized) residuals will be rejected, these are first arranged in descending order within the series. The probability statement corresponding to this situation is

$$\Pr(\xi_{y, \alpha/2} < y < \xi_{y, 1-\alpha/2}) = 1 - \alpha, \quad (13.13)$$

where  $y$  represents any of the statistics in TABLE 3, and  $\alpha$  denotes a new significance level, different from  $\alpha$ , which accounts for the simultaneity of all the tested elements of the series. A convenient way to handle this problem would be to derive a new probability density function for each  $y$  while using the customary significance level  $\alpha$ ; however, this is not always straightforward (for the derivation of such a  $\phi$  in row 4 of TABLE 3, see STEFANSKY [1972]). Here we follow the reasoning offered by THOMPSON [1935] and POPE [1976] for the general situation which leads instead to the modification of the significance level  $\alpha$ .

In the out-of-context approach, the selected significance level  $\alpha$  is related to the probability of  $y$  being in  $\langle \xi_{y,\alpha/2}, \xi_{y,1-\alpha/2} \rangle$  through (9). Realizing that all of the probability density functions of TABLE 3 are symmetrical, (9) can be rewritten as

$$1 - \alpha = \text{pr}(|y_i| < \xi_{y,1-\alpha/2}), \quad (13.14)$$

where  $\xi_{y,1-\alpha/2} = -\xi_{y,\alpha/2}$ . What needs be done is to consider all the  $y_i$ 's together, i.e., consider the probability of all the  $y_i$ 's being within  $\langle \xi_{y,\alpha/2}, \xi_{y,1-\alpha/2} \rangle$  simultaneously, or equivalently, the probability of the inequality  $|y_i| < \xi_{y,1-\alpha/2}$  holding simultaneously for all the  $y_i$ 's. Denoting this simultaneous probability by  $1 - a$ , one gets

$$1 - a = \text{pr}\left(\bigcap_{i=1}^N (|y_i| < \xi_{y,1-\alpha/2})\right). \quad (13.15)$$

If the  $y_i$ 's are statistically independent (rows 1 and 2 of TABLE 3), then the following equation is valid (see §3.4):

$$\text{pr}\left(\bigcap_{i=1}^N (|y_i| < \xi_{y,1-\alpha/2})\right) = \prod_{i=1}^N \text{pr}(|y_i| < \xi_{y,1-\alpha/2}) = \prod_{i=1}^N (1 - \alpha) = (1 - \alpha)^N, \quad (13.16)$$

and, thus, from (15)

$$1 - a = (1 - \alpha)^N. \quad (13.17)$$

It can be seen that the probability  $1 - a$  of all  $y_i$ 's being simultaneously inside  $\langle \xi_{y,\alpha/2}, \xi_{y,1-\alpha/2} \rangle$  is, theoretically, much smaller (for  $1 - \alpha < 1$ ) than the corresponding probability  $1 - \alpha$  for any single, isolated  $y_i$  taken out of context. Consequently, if the customary significance level  $\alpha$  is used when employing the in-context approach, then a smaller value must be used for the significance level for the individual components. From (17), for the confidence level  $1 - a$  one gets the corresponding in-context significance level

$$\alpha \doteq a/N, \quad (13.18)$$

and (15) can be rewritten as

$$\text{pr}\left(\bigcap_{i=1}^N (|y_i| < \xi_{y,1-\alpha/(2N)})\right) = \prod_{i=1}^N (1 - \alpha/N) \doteq 1 - \alpha. \quad (13.19)$$

Note that the in-context approach imposes stricter limits on the choice of  $\alpha$  since both  $1 - \alpha$  and  $1 - \alpha/(2N)$  must be positive.

The four in-context tests parallel the out-of-context tests in TABLE 3. Plotted in FIG. 10 are the *expansion factors*  $C_{0.05}$  needed to multiply the standard deviation  $s$  or  $\sigma$  to get the corresponding 95% confidence interval for both the out-of-context and the in-context tests. Clearly, the in-context tests use confidence intervals that are nearly twice as large as the out-of-context tests.

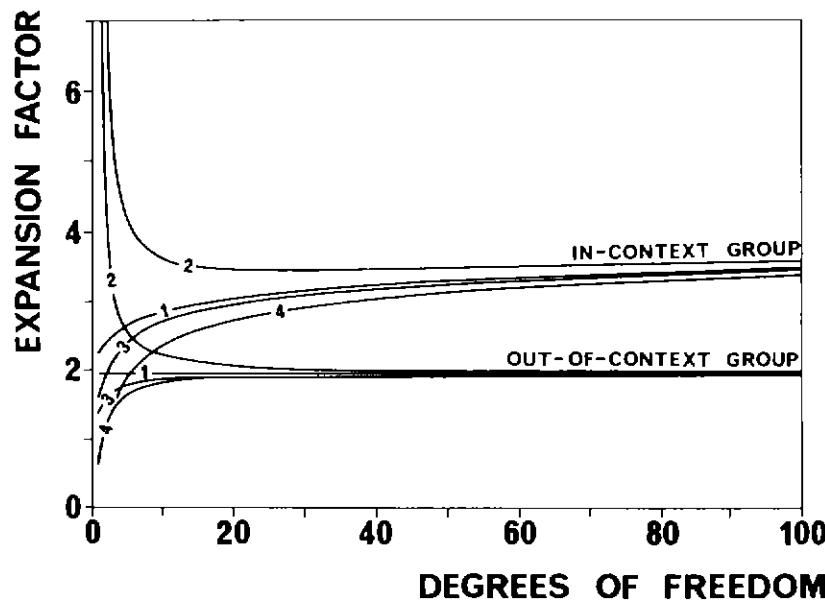


FIG. 13.10. Expansion factors  $C_{0.05}$  for confidence intervals for detection of outliers. (Numbering corresponds to the rows of Table 3.)

Because the last two statistics in TABLE 3 contain estimated residuals, which among themselves are not statistically independent, (15) is not exactly valid for them. The correlation (cf. §3.4) between any two estimated residuals is, however, fortunately small; namely (see (12)),

$$\rho = \frac{-1/N}{((N-1)/N)^{1/2}((N-1)/N)^{1/2}} = -\frac{1}{N-1}. \quad (13.20)$$

In any case, because the correlation is negative, the use of *Bonferroni's inequality* (e.g., MILLER [1966], FELLER [1968]),

$$\Pr\left(\bigcap_{i=1}^N (|y_i| < \xi_{y, 1-\alpha/(2N)})\right) \geq 1 - \sum_{i=1}^N \alpha/N = 1 - \alpha, \quad (13.21)$$

ascertains that the simultaneous probability of the estimated residuals is at least as large as that of the expected residuals, and the tests err on the side of caution. Generally, Bonferroni's inequality can be used when the need arises to sidestep the worrisome problem of what to do with statistical dependence that has been neglected.

#### 13.4. Simultaneous assessment of observations and mathematical models

Up to this point, only one row in the matrix of observations (10.16) has been examined at a time. There is no reason, however, why several or all elements of the observation vector cannot be analysed simultaneously. Such simultaneous analysis of all the elements of  $\mathbf{l} = [l_1, l_2, \dots, l_n]^T$  is referred to as joint multivariate analysis, or

simply *multivariate analysis*. Analysis of a subset of  $j$  elements ( $j < n$ ) is sometimes called *multivariate subset analysis*. The objective of the multivariate analysis is to determine how well the collection of observations fit the mathematical model (step (f) in §10.1).

To begin with, consider only two observations,  $\mathbf{l} = [l_1, l_2]^T$ , with a *bivariate normal probability density function* [HOGG AND CRAIG, 1970]:

$$\begin{aligned}\phi_l(\xi_1, \xi_2; \theta_1^{(l_1)}, \theta_2^{(l_1)}, \theta_1^{(l_2)}, \theta_2^{(l_2)}) &= n(\xi; \mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_{12}) \\ &= \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[ \left(\frac{\xi_1 - \mu_1}{\sigma_1}\right)^2 \right.\right. \\ &\quad \left.\left.- 2\rho\left(\frac{\xi_1 - \mu_1}{\sigma_1}\right)\left(\frac{\xi_2 - \mu_2}{\sigma_2}\right) + \left(\frac{\xi_2 - \mu_2}{\sigma_2}\right)^2\right]\right\}, \quad (13.22)\end{aligned}$$

where  $\mu_1$  and  $\mu_2$  denote the postulated means of the two observables associated with the variances  $\sigma_1^2, \sigma_2^2$  and the correlation coefficient  $\rho$ . Introducing

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}, \quad (13.23)$$

the bivariate normal probability density function becomes

$$\phi_l(\xi) = n(\xi; \boldsymbol{\mu}, \mathbf{C}) = \frac{1}{K} \exp\left[-\frac{1}{2} (\xi - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\xi - \boldsymbol{\mu})\right], \quad (13.24)$$

where (cf. §3.1)

$$K = 2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2} = 2\pi(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)^{1/2} = 2\pi(\det \mathbf{C})^{1/2},$$

$\boldsymbol{\mu}$  is the *bivariate population mean* ( $E(\mathbf{l}) = \boldsymbol{\mu}$ ), and  $\mathbf{C}$  is the *bivariate covariance matrix*. The expression for the probability density function is written so that it can be immediately generalized to any number  $n$  of observables simply by understanding  $\xi$ ,  $\boldsymbol{\mu}$ , and  $\mathbf{C}$  to be  $n$ -dimensional quantities. The constant  $K$  becomes [HAMILTON, 1967]  $K = (2\pi)^{n/2}(\det \mathbf{C})^{1/2}$ .

Let us now consider the world of experimentation. Least-squares solution (Chapter 12) provides only a least-squares estimate  $\hat{\mathbf{l}}$  of the expected value  $\bar{\mathbf{l}}$ . This means that the residual vector  $\mathbf{r}$  in the above probability density function must also be replaced by the estimate  $\hat{\mathbf{r}}$ , and so must be the covariance matrix  $\mathbf{C}$  and mean  $\boldsymbol{\mu}$ . Thus, instead of working with the normal density  $n(\xi; \boldsymbol{\mu}_l, \mathbf{C}_l)$ , or equivalently  $n(\xi; \boldsymbol{\mu}_{l-\mu} = \mathbf{0}, \mathbf{C}_l) = n(\xi; \boldsymbol{\mu}_r = \mathbf{0}, \mathbf{C}_r)$ , the only choice is to characterize the observations through the least-squares residuals  $\hat{\mathbf{r}}$  and their covariance matrix  $\mathbf{C}_{\hat{r}}$ . We get

$$\phi_{\hat{l}}(\xi) = n(\xi; \boldsymbol{\mu}_{\hat{l}}, \mathbf{C}_{\hat{l}}), \quad (13.25)$$

or equivalently

$$\phi_f(\xi) = n(\xi; \mu_{f-\hat{f}} = \mathbf{0}, C_{f-\hat{f}}). \quad (13.26)$$

We thus obtain

$$\phi_f(\xi) = \frac{1}{\hat{K}} \exp \left[ -\frac{1}{2} \xi^T C_f^{-1} \xi \right], \quad (13.27)$$

where

$$\hat{K} = (2\pi)^{n/2} (\det C_f)^{1/2}. \quad (13.28)$$

Note that in the case where  $\sigma_0^2$  is unknown, it is replaced by  $\hat{\sigma}_0^2$ , and  $C_f$  is replaced, in the above, by  $\hat{C}_f$  (cf. (12.56)). It should be pointed out that because  $C_f$  is always singular, the above probability density function has no meaning. This is why it is a normal practice to neglect the off-diagonal elements of  $C_f$  (or  $\hat{C}_f$ ) in dealing with  $\phi_f$ ; this aspect will be examined later.

A closer look at the vector  $\hat{r}$ , as well as a comparison of its  $\phi_f$  with the univariate probability density function, will now be helpful. Firstly, parallel to §13.3, it can be assumed that each constituent  $\hat{r}_i$  of  $\hat{r}$  is a random variable, and that there is no systematic effect in  $\hat{r}$ . Secondly, constituents  $\hat{r}_i$  are correlated whether or not the original observations  $l_i$  were statistically dependent to begin with; this correlation comes from the use of the mathematical model. The same situation has already been seen in the previous section, where the estimated residuals  $\hat{v}_i$  were correlated through the mathematical model. It is important to realize that when one speaks about the least-squares residual vector  $\hat{r}$ , one is referring to the observations and the mathematical model. Hence, the behaviour of  $\hat{r}$  is indicative of the behaviour of both  $l$  and the model  $f$ ; it is generally impossible to disentangle the two. This explains the reference to the simultaneous testing of the observations and the mathematical model in the title of this section.

It is interesting to note that the quadratic form in  $\phi_f$  is merely the norm, or distance from  $\mathbf{0}$ , used in the context of the least-squares method (cf. §12.1). When set equal to a constant, this quadratic form can be thought of as an equation of a hyperellipsoid in  $n$  dimensions (cf. §3.1). Use will be made of this very important property in the next section. Further,  $\mu$  is generally unknown except when  $\mu$  is expected to be equal to a zero vector; thus, the comments presented here will be mostly limited to the case of unknown  $\mu$ .

In *multivariate tests*, it is sometimes expedient to use quantities other than the residuals. If it is possible to determine by exactly how much the observation vector  $l$  fails to fit the mathematical model, e.g., if it is possible to compute a vector of disclosures, denoted here by  $w'$ , in the sense of (10.6), i.e.,  $w' = g(l)$ , then these disclosures are particularly suitable to use in assessing the observations. Even if it is

not possible to obtain  $w'$  for the whole vector  $I$ , it may be possible to do it for some subvectors of  $I$ . It should be pointed out that the misclosures  $w$ , resulting from linearization (12.3), are of a distinctly different kind; they are not suitable for testing because they depend on the arbitrarily selected points of expansion  $x^{(0)}, I^{(0)}$ , that do not lend themselves to testing.

The multivariate tests fall into two families. The first family consists of tests concerning the postulated probability density function and, thus, make simultaneous use of all the elements of the vector  $\hat{r}$  or  $w'$  (see TABLE 4). The second family is used to search for outliers and will be dealt with later. When designing multivariate tests for the observations, the first problem encountered is that every  $\hat{r}_i$  (or  $w'_i$ ) must be assumed to belong to a different population. In this context, it makes sense to talk about a probability density function of individual  $\hat{r}_i$ 's (or  $w'_i$ 's), because they are components of  $\hat{r}$  (or  $w'$ ) and, thus, random univariates. The difference between this case and the case discussed in §13.3, where  $r_i$  were individual values of a series, should be borne in mind. The vector  $\hat{r}$  (or  $w'$ ) thus has generally *statistically unhomogeneous* elements; evidence for this is contained in the covariance matrix  $C_{\hat{r}}$  (or  $C_{w'}$ ) the diagonal elements of which are generally different. This problem can be overcome by making the components homogeneous by standardizing them. Since all the components are assumed to belong to populations with different normal (or  $t$  or  $\tau$ ) density, their *standardization* is a straightforward task. It is accomplished by the well-known transformation (e.g., HOGG AND CRAIG [1970])

$$\tilde{l}_i = \frac{l_i - \mu_i}{\sigma_i}, \quad (13.29)$$

which is schematically displayed in FIG. 11. Since, as stated earlier, neither the actual mean  $\mu_i$ , nor the standard deviation  $\sigma_i$  is known, the transformation takes the form

$$\tilde{r}_i = \frac{l_i - \hat{l}_i}{\sigma_{l_i - \hat{l}_i}} = \frac{\hat{r}_i}{\sigma_{\hat{r}_i}}, \quad (13.30)$$

where  $\sigma_{\hat{r}_i}$  is the square root of the  $i$ th diagonal element of  $C_{\hat{r}}$ , and  $E(\hat{r}_i) = 0$ ; all covariances are neglected. The misclosures are transformed in a similar fashion: namely,  $\tilde{w}'_i = w'_i / \sigma_{w'_i}$ , where, again,  $E(w'_i) = 0$ . The probability density of each  $\tilde{r}_i$  (or  $\tilde{w}'_i$ ) is standard normal, i.e.,  $n(\xi; 0, 1)$ .

Above, it was tacitly assumed that the scale  $\sigma_0^2$  of the covariance matrix  $C_{\hat{r}}$  (or  $C_{w'}$ ) is known, i.e., that all the variances are properly scaled. When  $\sigma_0^2$  is unknown and replaced by  $\hat{\sigma}_0^2$ , the situation changes; the standardized residuals,

$$\tilde{r}_i = \frac{\hat{r}_i}{\hat{\sigma}_{\hat{r}_i}}, \quad (13.31)$$

have a  $\tau$  probability density function with degrees of freedom  $v$  equal to the

TABLE 13.4  
Assessment of observations and model

Name	Situation		$H_0$ (null hypothesis)		Statistic $y$	Probability density function of $y^a$	$1 - \alpha$ confidence interval for the quantity being tested	Remarks
	$\theta_1$	$\theta_2$	$\sigma_0^2$					
$\chi^2$ goodness of fit test	1 — known	$C_p$ or $C_w$ known	$\sigma_0^2$ known	histogram of standard- ized quantities	$\sum_{i=1}^n \frac{(a_i - e_i)^2}{e_i}$	$\chi^2(\xi; n-1)$	$0 < y < \xi_{\chi_m^{2, 1-\alpha}}$	$n = \text{number of classes.}$ $n-1 = \text{degrees of freedom.}$
of residuals or mis closures	2 — known	$\hat{C}_p$ or $\hat{C}_w$ used	$\sigma_0^2$ unknown $\hat{\sigma}_0^2$ used	compatible with $n(\xi; 0, 1)$ or $\tau(\xi; \nu)$ or $t(\xi; \nu)$	$\sum_{i=1}^n \frac{(a_i - e_i)^2}{e_i}$	$\chi^2(\xi; n-2)$	$0 < y < \xi_{\chi_m^{2, 1-\alpha}}$	one degree of freedom lost due to estimation of $\sigma_0^2$ .
Test of the quadratic form of the mis closures	3 — $\mu = 0$	$C_w$ known	$\sigma_0^2$ known	individual mis closures have the probability density function $n(\xi_i; 0, \sigma_{w_i}^2)$	$w'^T C_w^{-1} w'$	$\chi^2(\xi; m')$	$\xi_{\chi_m^{2, \alpha/2}} < y$ $< \xi_{\chi_{m-1}^{2, 1-\alpha/2}}$	$m' = \text{number of mis closures in}$ the subset; if $m' \leq m$ , this test is equivalent to the last test of this table.
	4 — $\mu = 0$	$\hat{C}_w$ used	$\sigma_0^2$ unknown $\hat{\sigma}_0^2$ used	individual mis closures have the probability density function $n(\xi_i; 0, \hat{\sigma}_{w_i}^2)$	$w'^T \hat{C}_w^{-1} w'$	$m' F(\xi; m', \nu)$	$m' \xi_{F_{m', \nu}, \alpha/2} < y$ $< m' \xi_{F_{m', \nu}, 1-\alpha/2}$	$\nu = m - u$ ; $\sigma_0^2$ and $w$ are statistically independent.
Test of the quadratic form of the residuals	5 — known	$C_p$ known	$\sigma_0^2$ known (tested)	residuals have the probability density function $n(\xi_i; 0, \sigma_{\epsilon_i}^2)$	$\nu \hat{\sigma}_0^2 / \sigma_0^2$	$\chi^2(\xi; \nu)$	$\frac{\nu \hat{\sigma}_0^2}{\xi_{\chi_{\nu, 1-\alpha/2}}^2} < \sigma_0^2$ $< \frac{\nu \hat{\sigma}_0^2}{\xi_{\chi_{\nu, \alpha/2}}^2}$	$\nu = m - u$ to test if $\sigma_0^2 = \hat{\sigma}_0^2$ .

<sup>a</sup> HAMILTON [1967]:  $\chi^2(\xi; \nu)$ —chi-squared density with  $\nu$  degrees of freedom.  
 $F(\xi; m', \nu)$ — $F$  density with  $m'$  and  $\nu$  degrees of freedom.

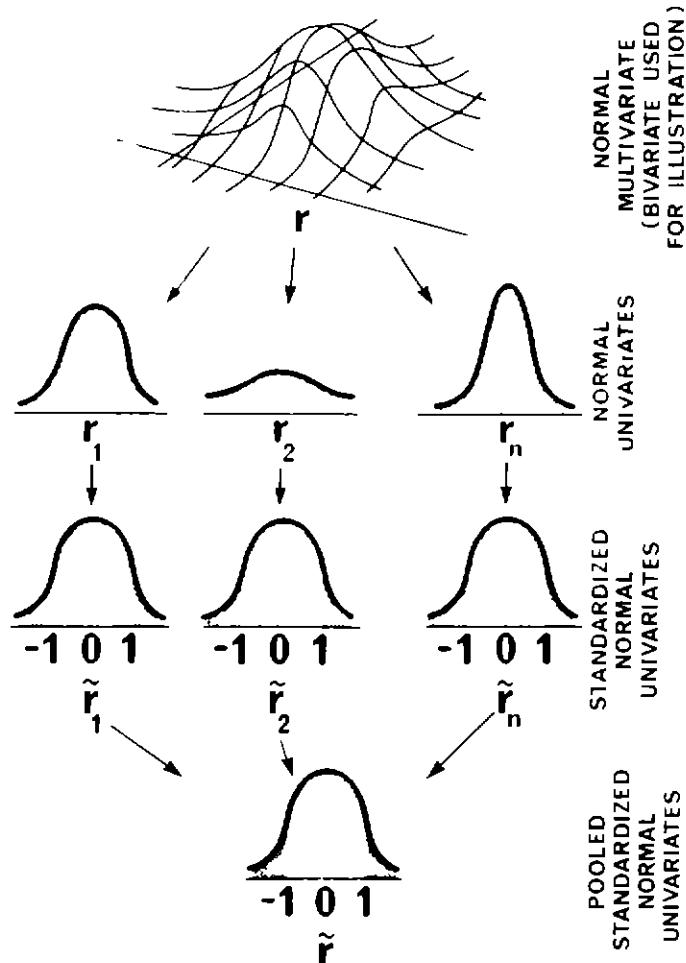


FIG. 13.11. Reduction of a multivariate to a univariate probability density function.

redundancy  $m - u$  in eqn. (12.51) used to obtain  $\hat{\sigma}_0^2$ . When pooled together, following a scheme parallel to that shown in Fig. 11, these individual  $\tau$  variates produce another  $\tau$  function with the same degrees of freedom. Standardization of the misclosures results in

$$\tilde{w}'_i = \frac{w'_i}{\hat{\sigma}_{w'_i}}, \quad (13.32)$$

distributed according to Student's  $t$  probability density function with  $v$  degrees of freedom; this is because  $\mu_i = 0$  is known. It is a requirement for Student's  $t$  function that the scale ( $\hat{\sigma}_0^2$ ) of the standard deviations  $\hat{\sigma}_{w'_i}$  be determined from some source independent of  $w'$ . When pooled together, the individual misclosures produce another  $t$  function with the same degrees of freedom.

It has to be pointed out that, in the above transformation of the multivariate to the univariate problem, the statistical dependence (covariances) among the  $\tilde{r}_i$ 's (or  $w'_i$ 's) was simply disregarded. As an alternative, one could diagonalize  $C_{\tilde{r}}$  (or  $C_{w'}$ )—see §3.1—, transform  $\tilde{r}$  (or  $w'$ ) into the so indicated eigenvector space to get a vector of  $m - u$  independent variates, and work with these independent variates. For

reasons of complexity and economy, this procedure is rarely followed. Instead, the Bonferroni inequality (21) approach is generally used.

The  $\chi^2$  goodness of fit test of residuals  $\hat{r}$  (or  $\hat{w}$ ) standardized in the described way (first test in TABLE 4) is then identical to that of the univariate case (TABLE 2). Another test that can be designed is the *test of the quadratic form of the misclosures*. Two distinctly different cases (rows 3 and 4, TABLE 4) occur:  $\sigma_0^2$  is either known—leading to the  $\chi^2$  statistic—or unknown—leading to the  $F$  statistic. It is interesting to note that when  $\sigma_0^2$  is known, the corresponding test can be performed for any subset of  $w''$ s before the total least-squares solution is made; this is clearly not possible when  $\sigma_0^2$  is unknown. The *test of the quadratic form of the residuals* (row 5, TABLE 4) is useful in testing the hypothesized value for  $\sigma_0^2$  in (12.42). The validity of this hypothesis can be tested using the knowledge that the statistic  $y = v\hat{\sigma}_0^2/\sigma_0^2$  has a  $\chi^2(\xi; v)$  probability density function. The test is also known as the  $\chi^2$  test of the variance factor.

These tests can fail for a number of reasons:

- (a) the non-normal density of the  $r$ 's (or  $w''$ s),
- (b) the incorrect mathematical model,
- (c) the presence of systematic errors in the observations, and
- (d) the incorrect a priori covariance matrix of the observations.

The last reason has been investigated in detail by MAGNESS AND MC GUIRE [1962]. A multivariate test of the quadratic form of the residuals, involving a second independent estimate of  $\sigma_0^2$ , can be found in HAMILTON [1967].

Let us turn our attention now to the second family of multivariate tests (TABLE 5) designed to search for outliers among the components of  $r$  and  $w'$ . There are two possible tests for the residuals  $r$  (rows 1 and 2) and two for the misclosures  $w'$  (rows 3 and 4, TABLE 5) depending on whether  $\sigma_0^2$  is known or unknown. They completely parallel the univariate tests of TABLE 3. Note that when testing each individual component  $r_i$  (or  $w'_i$ ) out of context, the covariances need not be considered, but the knowledge of the estimated variances is essential. These are, of course, extracted from  $\hat{C}_{\hat{r}}$  (or  $\hat{C}_{w'}$ ). POPE [1976] has investigated the general properties of the matrix  $\hat{C}_{\hat{r}}$  and has shown that an approximation of  $\hat{\sigma}_{\hat{r}}$  can be obtained even when  $\hat{C}_{\hat{r}}$  is not available. For the model explicit in I (10.8), he obtained

$$\hat{\sigma}_{\hat{r}_i} \doteq \left( \frac{n-u}{n} \right)^{1/2} \frac{\hat{\sigma}_0}{\sigma_0} \sigma_{r_i}. \quad (13.33)$$

Analogous equations can be derived for the other mathematical models.

To assess the components of the vector  $\hat{r}$  (or  $w'$ ) in context of the whole vector, the results developed in §13.3 are invoked. Simply, in TABLE 5,  $\alpha$  is replaced everywhere by  $a = \alpha/n$ , where  $n$  is the number of components of  $\hat{r}$  (or  $w'$ ). In tests applied in this manner, the covariances among the components of  $\hat{r}$  (or  $w'$ ) have again been disregarded. It is comforting to know that the absolute value of the covariances among the least-squares residuals  $\hat{v}$  should be small, as shown for a simple model in the univariate situation (20). When one deals with the statistically dependent residual vector  $s$  (§10.4), the problem of neglecting covariances can be more serious

TABLE 13.5  
Assessment of individual misdeceptions and residuals (out-of-context of the vector of values)

Name	Situation		$H_0$ (null hypothesis)	Statistic $\gamma$	Probability density function of $y^a$	$1-\alpha$ confidence interval for the quantity being tested <sup>b</sup>	Remarks
	$\theta_1$	$\theta_2$					
Test of a residual outlier	1	$C_r$ known	$\sigma_0^2$	$\hat{r}_i$ belongs to a sample having the probability density function $n(\xi_i; 0, \sigma_{\hat{r}_i}^2)$	$\hat{r}_i = \frac{\hat{r}_i}{\sigma_{\hat{r}_i}}$	$n(\xi; 0, 1)$	$-\xi_{n(0, 1), 1-\alpha/2}\hat{\theta}_{\hat{r}_i} < \hat{r}_i < \xi_{n(0, 1), 1-\alpha/2}\sigma_{\hat{r}_i}$
	2	$\mu = 0$ known	$\sigma_0^2$	$\hat{r}_i$ belongs to a sample having the probability density function $n(\xi_i; 0, \hat{\sigma}_{\hat{r}_i}^2)$	$\hat{r}_i = \frac{\hat{r}_i}{\hat{\sigma}_{\hat{r}_i}}$	$\tau(\xi; \nu)$	$-\xi_{\tau, 1-\alpha/2}\hat{\theta}_{\hat{r}_i} < \hat{r}_i < \xi_{\tau, 1-\alpha/2}\hat{\sigma}_{\hat{r}_i}$
Test of a disclosure outlier	3	$C_{w'}$ known	$\sigma_0^2$	$w'_i$ belongs to a sample having the probability density function $n(\xi_i; 0, \sigma_{w'_i}^2)$	$\hat{w}'_i = \frac{w'_i}{\sigma_{w'_i}}$	$n(\xi; 0, 1)$	$-\xi_{n(0, 1), 1-\alpha/2}\hat{\sigma}_{w'_i} < w'_i < \xi_{n(0, 1), 1-\alpha/2}\sigma_{w'_i}$
	4	$\mu = 0$ known	$\sigma_0^2$	$w'_i$ belongs to a sample having the probability density function $n(\xi_i; 0, \hat{\sigma}_{w'_i}^2)$	$\hat{w}'_i = \frac{w'_i}{\hat{\sigma}_{w'_i}}$ or $\tau(\xi; \nu)$	$t(\xi; \nu)$ or $\tau(\xi; \nu)$	$-\xi_{\tau, 1-\alpha/2}\hat{\sigma}_{w'_i} < w'_i < \xi_{\tau, 1-\alpha/2}\sigma_{w'_i}$

<sup>a</sup>POPE [1976]:  $n(\xi; 0, 1)$ —standard normal density with a mean of 0 and a variance of 1.  
 $t(\xi; \nu) - t$ —density with  $\nu$  degrees of freedom.

<sup>b</sup>For in-context testing of vectors  $w'$  and  $r$ : Replace  $\alpha$  with  $\alpha/n$ , where  $n$  is the dimension of  $w'$  or  $r$ .

because there may be a significant statistical dependence among the components even before the adjustment takes place. In such a situation, the covariance matrix of  $\hat{s}$  should be either diagonalized or Bonferroni's inequality employed.

### 13.5. Assessment of the determined parameters

After the observations have been screened and the mathematical model examined, an assessment of the parameters  $x$  (or equivalently  $\lambda$ ), determined through the least-squares process, should take place (cf. step (g) in §10.1). The assessment consists of the establishment of confidence regions for the parameters, which represent the amount of trust that can be placed on the estimated values of  $x$  (or  $\lambda$ ). Examination of the compatibility of two independent determinations of the same unknown parameters may become part of the assessment, if more than one independent determination has been done.

The ultimate goal set up in §13.1 was to characterize the uncertainties in  $\hat{x}$  through probability, because probability gives a direct measure of the trust one can have in the results. Accordingly, we should now be concerned with the multivariate probability density function associated with the parameters, because it describes the probability density in the  $u$ -dimensional parameter space. The probability density function of the observations was postulated to be the multivariate normal given by (25). As the parameters  $x$  are, at least in the neighbourhood of the point of expansion, simply linear combinations of the observations  $t$  (cf. FIG. 12.1), they can also be regarded as stochastical quantities and it is known (e.g., HAMILTON [1967]) that their probability density function will also be multivariate normal. Thus, in equation form (cf. (24)):

$$\phi_x(\xi) = n(\xi; \mu_x, C_x) = \frac{1}{K} \exp\left[-\frac{1}{2}(\xi - \mu_x)^T C_x^{-1}(\xi - \mu_x)\right], \quad (13.34)$$

where

$$K = (2\pi)^{u/2} (\det C_x)^{1/2}.$$

In the above,  $\mu_x$  is the vector of expected values of the parameters, and  $C_x$  is the expected covariance matrix of the parameters. In reality, only the estimate  $\hat{x}$  of  $\mu_x$  and  $C_{\hat{x}}$  (or  $\hat{C}_{\hat{x}}$ ) of  $C_x$  can be obtained. Thus the probability density function must be postulated as

$$\phi_x(\xi) = n(\xi; \hat{x}, C_{\hat{x}}) = \frac{1}{\hat{K}} \exp\left[-\frac{1}{2}(\xi - \hat{x})^T C_{\hat{x}}^{-1}(\xi - \hat{x})\right], \quad (13.35)$$

where

$$\hat{K} = (2\pi)^{u/2} (\det C_{\hat{x}})^{1/2}.$$

The sought measure of trust in  $\hat{x}$  is then obtained through a  $u$ -dimensional integration over the neighbourhood of  $\hat{x}$ , much the same way as for the unidimensional case (§13.1).

The two tests shown herein provide practical alternatives to the measurement of the degree of trust in  $\hat{x}$ . They correspond to the cases when  $\sigma_0^2$  is either known or unknown.

(a) Let us consider first that  $\sigma_0^2$  is known; i.e.,  $C_l$  is known and, thus,  $C_{\hat{x}}$  is available. The statistic used is then

$$y = (\mathbf{x} - \hat{\mathbf{x}})^T C_{\hat{x}}^{-1} (\mathbf{x} - \hat{\mathbf{x}}). \quad (13.36)$$

For  $\mathbf{x}$  and  $\hat{\mathbf{x}}$  distributed normally, the probability density function of  $y$  is  $\chi^2(\xi; u)$  [GRAYBILL, 1976]. Now, for a specific significance level  $\alpha$ , one gets a specific value of  $y$  that will, again, be denoted here by  $\xi_{y,1-\alpha}$ . Substitution of this value into (36) converts the equation into an equation of a  $u$ -dimensional hyperellipsoid, a fact already alluded to in the previous section. This hyperellipsoid can be understood as a  $u$ -dimensional confidence region centred on  $\hat{\mathbf{x}}$ . Any tested value  $\mathbf{x}$  that falls within the hyperellipsoid must then be considered compatible with  $\hat{\mathbf{x}}$  on the level of probability  $1 - \alpha$ . The concept of hyperellipsoids will be used extensively in Part IV.

(b) The case of unknown  $\sigma_0^2$  is best tested using the following statistic:

$$y = \frac{(\mathbf{x} - \hat{\mathbf{x}})^T C_{\hat{x}}^{-1} (\mathbf{x} - \hat{\mathbf{x}})/u}{[(m-u)\hat{\sigma}_0^2/\sigma_0^2]/(m-u)} = \frac{(\mathbf{x} - \hat{\mathbf{x}})^T \hat{C}_{\hat{x}}^{-1} (\mathbf{x} - \hat{\mathbf{x}})}{u}. \quad (13.37)$$

It has an  $F(\xi; u, v)$  probability density function [HOGG AND CRAIG, 1970]. This statistic is used again to check the compatibility between a hypothesized set of values for  $\mathbf{x}$  and those for  $\hat{\mathbf{x}}$ , in the same manner as above. Both of the above tests are rigorous; the way the statistics are formulated, the tests take into account all the variances and covariances.

One runs into difficulty, however, when subvectors of the determined parameters are examined separately, something often done in practice. To investigate this situation, let us concentrate on case (a), with the second case being analogous, and write the quadratic form in (36) as  $\Delta \mathbf{x}^T C_{\hat{x}}^{-1} \Delta \mathbf{x}$ . Denoting a subvector of  $\hat{\mathbf{x}}$  by a subscript, say  $k$ , and its corresponding quadratic form by  $\Delta \mathbf{x}_k^T C_{\hat{x}_k}^{-1} \Delta \mathbf{x}_k$ , where  $C_{\hat{x}_k}^{-1}$  is the appropriate submatrix of  $C_{\hat{x}}^{-1}$ , the probability statement for such a subset can be written as

$$\text{pr}(\Delta \mathbf{x}_k^T C_{\hat{x}_k}^{-1} \Delta \mathbf{x}_k \leq \xi_{y,1-\alpha}) = 1 - \alpha, \quad (13.38)$$

where  $\xi_{y,1-\alpha}$  is determined from the  $\chi^2(\xi; q)$  in which  $q < u$  is equal to  $\dim(\hat{\mathbf{x}}_k)$ . Now, the simultaneous probability of the quadratic forms of all the  $N$  separate subvectors is smaller so, going through the same argument as in §13.4, one obtains the applicable Bonferroni's inequality as

$$\text{pr}\left(\bigcap_{k=1}^N (\Delta \mathbf{x}_k^T C_{\hat{x}_k}^{-1} \Delta \mathbf{x}_k \leq \xi_{y,1-\alpha})\right) \geq 1 - \sum_{k=1}^N \alpha = 1 - N\alpha. \quad (13.39)$$

It shows that if the probability  $1 - \alpha$  is required for the testing of the subvector in-context of the complete solution, then the individual confidence regions of the subvectors must be increased so as to reflect the change from  $\xi_{y,1-\alpha}$  to  $\xi_{y,1-\alpha/u}$ . This, clearly, limits the choice of  $\alpha$  because, again,  $1 - \alpha$  and  $1 - N\alpha$  must be positive.

It is instructive to examine the inequality (39) more closely. If one considers the special case of  $q = 1$ , i.e., the case of the subvectors of one element, or the individual components of  $x$  (or  $\lambda$ ) and  $N = u$ , the simultaneous probability statement can be written as

$$\text{pr} \left( \bigcap_{k=1}^u \left( \Delta x_k^2 / \sigma_{\hat{x}_k}^2 \leq \xi_{y,1-\alpha/u} \right) \right) \geq 1 - \alpha,$$

or

$$\text{pr} \left( \bigcap_{k=1}^u \left( \Delta x_k / \sigma_{\hat{x}_k} \leq \sqrt{\xi_{y,1-\alpha/u}} \right) \right) \geq 1 - \alpha, \quad (13.40)$$

where  $y$  has the  $\chi^2(\xi; 1)$  probability density function. It can be seen that the square root of the abscissa value, i.e.,  $\sqrt{\xi_{y,1-\alpha/u}} = C_\alpha(u)$ , is the expansion factor needed to multiply the standard confidence interval, e.g.,  $\langle \hat{x}_k - \sigma_{\hat{x}_k}, \hat{x}_k + \sigma_{\hat{x}_k} \rangle$ , to obtain the  $1 - \alpha$  in-context confidence interval:

$$|x_k - \hat{x}_k| \leq C_\alpha(u) \sigma_{\hat{x}_k}. \quad (13.41)$$

Plotted in FIG. 12 are the expansion factors  $C_{0.05}$  for  $q = 1$  and various values of  $u$ , as functions of the degrees of freedom  $v = m - u$  of the adjustment. It is seen that for growing degrees of freedom  $v$ , the plotted values of  $C_\alpha$ , based on the  $F$  density ( $\sigma_0^2$  unknown), tend toward the corresponding values based on the  $\chi^2$  density ( $\sigma_0^2$  known). It is interesting to compare the present figure with FIG. 10. In FIG. 12, the curve for  $u = 1$  corresponds exactly to the out-of-context curve denoted by 2 in FIG. 10.

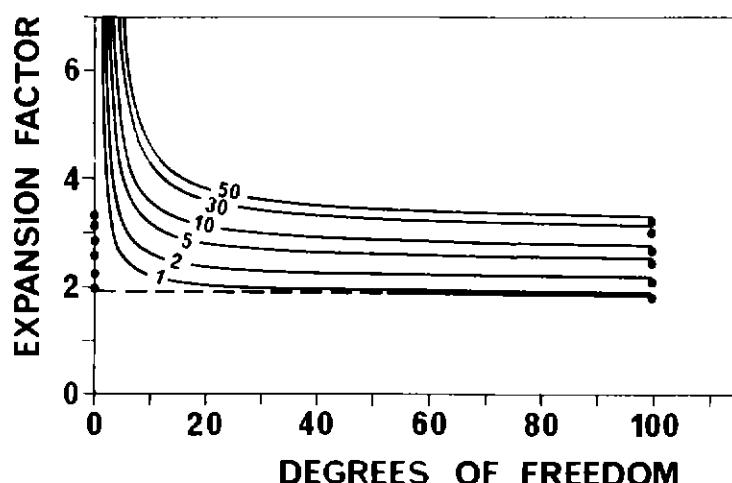


FIG. 13.12. Expansion factors  $C_{0.05}$  for confidence intervals of results. • denote  $\sigma_0^2$  known, — denote  $\sigma_0^2$  unknown, values of  $u$  indicated on curves.

## CHAPTER 14

### FORMULATION AND SOLVING OF PROBLEMS

This chapter deals with the types of problems most widely encountered in geodesy. In the opening section, the problem of design of a geodetic project, i.e., the design of how to achieve desired results with the minimum accuracy of observations, is treated.

The decomposition of an observable (10.41) was selected as the most natural key to the classification of the remaining standard problems. The presence or absence of the individual terms in the decomposition formula leads to various models suitable for solving certain problems. There are 16 possible cases out of which 15 make sense: the case of  $I = \mathbf{0}$  does not. Of the remaining 15, the three situations when  $v = s = \mathbf{0}$  will not be treated because they evidently correspond to cases with consistent equations. Such models have already been fully treated in §11.2. Further, the cases of observed statistically independent ( $I = v$ ) and statistically dependent random noise ( $I = s$ ) do not present too much of a problem. The only meaningful questions that could be posed in this case would concern the statistical aspects of the samples  $v$  or  $s$ . These tasks have already been adequately treated in Chapters 10 and 13, thus they will be left out here. This still leaves 10 cases to deal with, and these are listed in TABLE 1.

Generally, the presence of the term  $\Phi^T \lambda$  typifies the regression problem which is treated, together with some related problems, in the second section of this chapter. The presence of the term  $Hx$  (or  $Ax$ ) is said to constitute the problem of adjustment, which is treated in the third section. The simultaneous presence of  $v$  and  $s$  in the model distinguishes a subclass of problems with two random components. This is also dealt with in the third section.

The remainder of this chapter deals with complications encountered in the formulation and solving of problems. The fourth section covers the complications arising from the introduction of a priori—objective or subjective—statistical information about the unknown parameters. In the fifth section, problems complicated by constraints and singularities are addressed. The last section describes the techniques for solving for the unknown parameters when the observations are not all available at the same time or when the unknown parameters change with time.

TABLE 14.1

Standard problems. (1 denotes presence of the term, 0 denotes absence of the term.)

Name (label)	$Hx$	$\Phi^\top \lambda$	$s$	$v$	Section Treated
Simple regression	0	1	0	1	
Regression	0	1	1	0	14.2
Two-component regression	0	1	1	1	
Simple adjustment (of observations)	1	0	0	1	
Adjustment (of observations)	1	0	1	0	
Two-component adjustment (of observations)	1	0	1	1	
Simple simultaneous adjustment and regression	1	1	0	1	14.3
Simultaneous adjustment and regression	1	1	1	0	
Two-component simultaneous regression and adjustment	1	1	1	1	
Random series decomposition	0	0	1	1	

### 14.1. Optimal accuracy design

The main objective of design for optimal accuracy, or simply preanalysis, is to derive specifications for the necessary and sufficient accuracy of a proposed set of observables given the desired accuracy of the unknown parameters (cf. step (c) in §10.1). Preanalysis is, of course, carried out before the measurements are actually made. The benefits accruing from preanalysis are obvious and need not be articulated.

There are three particular cases of the general problem of preanalysis:

(a) In this case, called the *first-order design problem*, the covariance matrices  $C_{\hat{x}}, C_t$  are given, and the determination of the design matrix  $G$  (see (10.4)), or  $A$ , or  $A, B$  (see (10.11)), is required.

(b) In this case, called the *second-order design problem*,  $C_{\hat{x}}$  and  $G$  (or  $A$ , or  $A, B$ ) are given, and the covariance matrix  $C_t$  of the observations is required.

(c) In this case, called the *combined design problem*,  $C_{\hat{x}}$  is given, and both the design matrix and covariance matrix of the observations are required.

Essentially, there are two approaches that can be used to solve these design problems—a direct approach and a trial and error procedure. Our discussion of the direct approach will be restricted to case (b); any solution to the other two problems requires special mathematical techniques, namely, linear and quadratic programming, considered to be outside the scope of this book. The interested reader is referred to the literature (e.g., GRAFARENDS AND HARLAND [1973], GRAFARENDS [1974], and CROSS AND THAPA [1979]). On the other hand, the trial and error procedure can be used to solve any of the three cases.

Let us begin with the direct method of solving for the covariance matrix of the observations. For simplicity, consider the model explicit in  $t$  (10.9). Application of

the covariance law yields, realizing that  $w$  is constant,

$$\hat{\mathbf{C}}_l = \mathbf{H} \hat{\mathbf{C}}_{\hat{x}} \mathbf{H}^T, \quad (14.1)$$

from which the covariance matrix of the observations can be computed. A subtle point, usually ignored in the existing literature, must be made regarding  $\hat{\mathbf{C}}_l$ : strictly speaking, this is the covariance matrix of the adjusted observations given by (12.41). Making measurements to this accuracy would be too stringent a requirement, because the covariances of the observations acquired in the field are somewhat larger. Using (12.41) and (12.38) with  $\mathbf{B} = -\mathbf{I}$ , we can write

$$\mathbf{C}_l - \mathbf{C}_l \mathbf{L} \mathbf{C}_l = \mathbf{H} \hat{\mathbf{C}}_{\hat{x}} \mathbf{H}^T, \quad (14.2)$$

i.e., an equation of second order in  $\mathbf{C}_l$ . Since the second-order term is much smaller than the linear term, the equation can be solved for  $\mathbf{C}_l$  through iterations; this is, however, usually not done.

It is interesting to point out the properties of the so-derived covariance matrix  $\hat{\mathbf{C}}_l$ : it is fully populated, square, symmetric, and, for  $n > u$ , singular. The latter characteristic originates with the fact that a  $u$ -dimensional parameter space is expanded into an  $n$ -dimensional observation space through  $\mathbf{H}$ . In terms of the weight matrix of the adjusted observations, we can get, using the pseudo-inverse introduced in §11.3,

$$\hat{\mathbf{P}}_l = \hat{\mathbf{C}}_l^\ddagger = (\mathbf{H} \hat{\mathbf{C}}_{\hat{x}} \mathbf{H}^T)^+ = (\mathbf{H}^T)^+ \hat{\mathbf{C}}_{\hat{x}}^{-1} \mathbf{H}^+. \quad (14.3)$$

BOSSLER ET AL. [1973], who have derived this expression by a different method, have also shown that, in general, there does not exist any inverse giving a diagonal, positive, definite weight matrix  $\hat{\mathbf{P}}_l$ . Under certain conditions, however, it is possible to obtain at least an approximate solution for a diagonal  $\hat{\mathbf{P}}_l$ ; we shall show here a method outlined by GRAFarend [1974] and worked out by STEEVES [1978]. To begin with, let us rewrite the equation for the covariance matrix of the parameters (12.36) as follows:

$$\hat{\mathbf{C}}_{\hat{x}} = \mathbf{N}^{-1} = (\mathbf{H}^T \mathbf{P}_l \mathbf{H})^{-1}. \quad (14.4)$$

From the above it is possible to write  $r$ -independent linear equations for the elements  $P_{ki}$  of the  $\mathbf{P}$  matrix: namely,

$$\sum_{k=1}^n h_{ki} \sum_{t=1}^n h_{tj} P_{kt} - n_{ij} = 0; \quad i, j = 1, \dots, u; \quad i \leq j, \quad (14.5)$$

in terms of the elements  $h_{kj}$  of the design matrix and elements  $n_{ij}$  of  $\mathbf{N} = \hat{\mathbf{C}}_{\hat{x}}^{-1}$ . By denoting the vector of  $r = \frac{1}{2}u(u+1)$  upper triangular elements of  $\mathbf{N}$  by  $\tilde{\mathbf{N}}$ ; the vector of  $q \leq r$  unknown elements of  $\mathbf{P}_l$  by  $\tilde{\mathbf{P}}$ ; and the corresponding  $r$  by  $q$  matrix composed of products of coefficients  $h_{ij}$  by  $\mathbf{E}$ , we get [MIKHAIL, 1976]

$$\tilde{\mathbf{P}} = (\mathbf{E}^T \mathbf{E})^{-1} \mathbf{E}^T \tilde{\mathbf{N}}. \quad (14.6)$$

Note that this approach yields directly the inverse of the covariance matrix of the observations. Thus the difficulty with the inversion of the covariance matrix of the adjusted observations is eliminated. There is, however, another difficulty inherent in the realistic prescription of covariances in  $\mathbf{C}_{\hat{x}}$ . For example, it is not sufficient to stipulate just a diagonal  $\mathbf{C}_{\hat{x}}$  matrix. More research needs be done in this domain.

In the trial and error approach, various design matrices or covariance matrices or both of them ( $\mathbf{H}, \mathbf{C}_l$ ) are tried separately or simultaneously until a satisfactory result ( $\mathbf{C}_{\hat{x}}$ ) is obtained. To illustrate this procedure, consider, again, the second-order design problem, i.e.,  $\mathbf{C}_{\hat{x}}$  and  $\mathbf{H}$  as given, and  $\mathbf{C}_l$  as sought. A certain  $\mathbf{C}_l^{(0)}$  is selected, and the corresponding value of  $\mathbf{C}_{\hat{x}}^{(0)}$  is computed from (12.36). This process is repeated  $(t - 1)$  times until the required  $\mathbf{C}_l^{(t)}$  produces a  $\mathbf{C}_{\hat{x}}^{(t)}$  that is close enough to the one desired. Clearly,  $\mathbf{H}$  can be changed simultaneously with  $\mathbf{C}_l$  if the combined design problem is faced. The main difficulty with this technique is the number of attempts it requires before satisfactory results are obtained; the convergence of the iterations to the best solution is difficult to achieve since there is no objective criterion to deploy. A tool that can help to overcome this obstacle is that of iterative computer graphics [NICKERSON ET AL., 1978]. It is clear that the results obtained by this technique, although usually acceptable in practice, are not strictly optimal.

## 14.2. Analysis of trend

The problem of *analysis of trend* arises when a data series is investigated (cf. §10.3). A *data series* may be a set of repeated observations  $l(\tau_i)$  of the same quantity in time, or it may simply be a set of observations distributed in space, with  $\tau$  being a space coordinate. The most important characteristic of the analysis of trend is that it uses a certain (time or space) sequence of the collected observations (data). This means that the trend in the matrix of observations (cf. (10.16)) in either the row sense, i.e., the  $l(\tau)$ , or in the column sense, i.e., the vector  $\mathbf{l}$ , can be analysed. It is customary to first analyse the row trend and then, once the data have been introduced into the main mathematical model and a solution made, analyse the trend of, say, the residual vector  $\hat{\mathbf{r}}$  to see whether any systematic effects remain.

The first step in investigating the behaviour of a series of observations  $l(\tau_i)$  is to decompose the series (FIG. 1) into the trend  $t(\tau_i)$  and residual  $r(\tau_i)$  components. Generally, this decomposition can be carried out only if the analytical shape (model) of the trend is known or stipulated. Let us suppose, then, that the analytical shape of  $t$  is known. This means that the systematic behaviour of the observable is known with respect to the coordinate  $\tau$ . Without any loss of generality, the model can then be written as a *linear form* (generalized polynomial) (cf. (10.17) and (10.19)):

$$t(\tau_i) = l(\tau_i) + r(\tau_i) = \sum_{j=1}^u \lambda_j \phi_j(\tau_i), \quad (14.7)$$

where  $[\phi_1, \phi_2, \dots, \phi_u] = \tilde{\phi}^T$  is a vector of functions of  $\tau$ , and  $[\lambda_1, \lambda_2, \dots, \lambda_u] = \lambda^T$  is a vector of constant coefficients that have to be determined. When considering  $N$

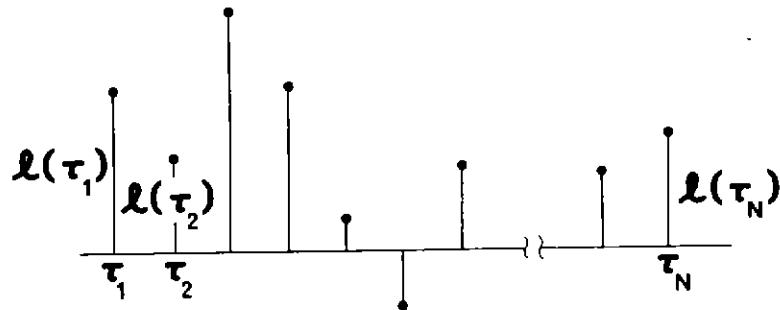


FIG. 14.1. A series of observations.

observables together,  $l(\tau_i)$  becomes  $l$  and we get

$$t = l + r = \Phi^T(\mathcal{T})\lambda, \quad (14.8)$$

where the scalar product is carried out in the  $u$ -dimensional space and is a function of  $\tau$ . The matrix  $\Phi^T(\mathcal{T})$  is, of course, the Vandermonde matrix (see §3.1). The linear form  $\tilde{\Phi}^T(\tau_i)\lambda$ , where  $\tilde{\Phi}(\tau_i)$  is the  $i$ th column of  $\Phi$ , is usually called the *approximant* to  $t(\tau_i)$ .

Notice that the Vandermonde matrix is constant once the sampling space  $\mathcal{T} \equiv \{\tau_1, \tau_2, \dots, \tau_N\}$  is given and the functions  $\phi_i$  selected. Thus, if the vector of coefficients is regarded as a special case of unknown parameters  $x$ , and the Vandermonde matrix as a special case of a design matrix  $H$ , then the observations can be written as follows:

$$l = Hx - r, \quad (14.9)$$

or

$$r = Hx + w, \quad w = -l. \quad (14.10)$$

This is merely the model explicit in  $l$  discussed earlier. Having made this connection between the trend analysis and models, let us first introduce the problem of data series interpolation.

If the number of selected functions ( $u$ ) equals the number of sampling points ( $N$ ), then the Vandermonde matrix is square and we are in the realm of *interpolation*. If, in addition, the functions  $\phi_i$  are linearly independent—such functions are usually called *base functions*—on  $\mathcal{T}$ , then the Vandermonde matrix is regular and its inversion exists. In this case, the solution (i.e., the vector of coefficients) is given by

$$\lambda = (\Phi^T(\mathcal{T}))^{-1} t.$$

(14.11)

This solution is known as interpolation of the series  $t = l + r$ . The literature provides innumerable schemes for solving the above equation (see, e.g., DAVIS [1963]). Interpolation thus reproduces the values of  $t$  at the sampling points  $\mathcal{T}$  but tells us nothing about  $r$ . It may be argued that interpolation yields a particular solution for  $r$ , specifically the trivial solution  $r = 0$ , regardless of how the properties of  $r$  are

specified. However, this viewpoint is difficult to substantiate when  $\mathbf{l}$  are observed values and the elements  $\mathbf{r}$  are interpreted as errors (noise) in the observations and, as such, are expected to be different from zero. For this reason, interpolation is seldom used in geodesy.

If the number of base functions is larger than the number of sampling points, i.e.,  $u > N$ , the problem is underdetermined; it can be reduced to an interpolation problem using the technique described in §11.3. Of much more interest is the case when  $u < N$ ; this problem is referred to as the *problem of approximation* [DAVIS, 1963; DRAPER AND SMITH, 1967].

The possibilities of obtaining a unique  $\lambda$  if the model (8) is overdetermined were discussed in §11.4. As already shown, depending on the metric space chosen, we may have the *uniform* (Tchebychev's) *approximation*, the *mean quadratic* (least-squares) *approximation*, or some other. Further, there are additional possibilities based on different concepts. *Spline approximation* [AHLBERG ET AL., 1967] should at least be mentioned here, and more ideas can be found in the literature (e.g., DAVIS [1963], CHENEY [1966]).

Among the scores of techniques, geodesy favours the least-squares approximation, which will be called *regression* here. Regression yields the solution  $\hat{\lambda}$  that minimizes the noise  $\mathbf{r}$  in the observation space  $\mathcal{L}$  metricized with least-squares metric. The least-squares technique has been developed in sufficient detail in Chapter 12 and can be used without any modification here. Specifically, if the observation space  $\mathcal{L}$  (spanned by  $\mathbf{l}$ ) is metricized by  $\mathbf{C}_r^{-1}$  (either diagonal or fully populated),  $\mathbf{C}_r^{-1}$  can express either the preconceived (desired) degrees of goodness of approximation for various values of  $\tau_i$  or the a priori knowledge of the goodness of observations  $\mathbf{l}$ . The method is called *simple regression*, if  $\mathbf{C}_r$  is diagonal, or just regression if it is not (TABLE 1). If we know, or are willing to postulate, the statistical properties of  $\mathbf{v}$  and  $\mathbf{s}$  in terms of  $\mathbf{C}_v$  and  $\mathbf{C}_s$  simultaneously, we get into the area of *two-component regression*, a method which will be discussed in §14.3.

One topic of particular interest here is least-squares regression using *orthogonal bases*. In Chapter 12, the system of least-squares normal equations was derived (12.21); an identical system holds for  $\hat{\lambda}$ ; namely,

$$(\Phi(\mathcal{T})\mathbf{C}_r^{-1}\Phi^T(\mathcal{T}))\hat{\lambda} = \Phi(\mathcal{T})\mathbf{C}_r^{-1}\mathbf{l} = -\mathbf{u}. \quad (14.12)$$

In this context, the matrix of normal equations is called *Gram's matrix*. It is written as

$$\tilde{\mathbf{G}} = \Phi(\mathcal{T})\mathbf{C}_r^{-1}\Phi^T(\mathcal{T}), \quad (14.13)$$

and is merely a matrix of the scalar products of all the possible pairs of vectors of base functional values  $[\phi_i(\tau_1), \phi_i(\tau_2), \dots, \phi_i(\tau_N)]$  in the space metricized by  $\mathbf{C}_r^{-1}$ . The diagonal elements are scalar products of the base functions with themselves, i.e., the squares of the magnitudes of the vectors of functional values. It is left to the reader to satisfy himself that these elements are also squares of norms of the appropriate functions in the sense of eqn. (11.9). Evidently, systems of mutually orthogonal base

functions (see §3.2) produce normal equations that are of a diagonal form: all off-diagonal elements are equal to zero. In addition, diagonal elements  $g_{ii}$  in  $\tilde{\mathbf{G}}$  are all positive. The advantage of such a system is that the equations are no longer interdependent and can thus be solved individually. Therefore we have

$$\hat{\lambda}_i = -u_i/g_{ii}, \quad i=1,\dots,u. \quad (14.14)$$

Any orthogonal system is orthogonal only for a special distribution of sampling points and many can be found in, e.g., ABRAMOWITZ AND STEGUN [1964]. Also, any system of base functions can be transformed to an orthogonal system through one of many orthogonalization processes of which the *Gram–Schmidt process* is the best known (see, e.g., CHENEY [1966]). An orthogonal system of functions for which the Gram matrix is an identity matrix is known as orthonormal. Any orthogonal base can be converted to an *orthonormal base* by dividing the functions  $\phi_i$  by  $(g_{ii})^{1/2} = \|\phi_i\|$ .

Finally, a few words should be directed to the selection of the base functions  $\phi$ . As already mentioned in §10.4, the base functions should be selected in accordance with our understanding of the observables and their role in the measuring process. In some instances, the base functions may be numerical functions obtained through the measurement of other natural phenomena affecting the observable or the measuring apparatus—for an example see §19.1. In other instances, the selection may reflect the behaviour of  $l$  as predicted by a law of physics or geometry—for an example see §20.2. If no such approach is possible, the base functions have to be selected arbitrarily—for an example see §27.4. More often than not, when we use the least-squares regression we are not at all sure that the model, i.e., the selected base, is justified. In this case, the residuals  $r(\tau_i)$  or  $r$  must be regarded as expressing not only the uncertainties in  $l$ , as we have so far always assumed, but also the uncertainties in the model. The extreme case occurs when the series  $l(\tau_i)$  is not an outcome of a measurement at all, and  $r(\tau_i)$  must be regarded as characterizing solely the misfit of the model to  $l(\tau_i)$ . Any statistical testing of  $r$  must therefore be performed with the above in mind. Further, it is often desirable to test whether or not a determined parameter  $\lambda_i$  is statistically significant, because it is not sure that  $\phi_i$  in the model should have been considered in the first place. Such testing can be performed using (13.41), where  $\hat{\lambda}$  simply replaces  $\hat{x}$  and  $\mathbf{0}$  replaces  $x$ . The test then examines whether the particular  $\hat{\lambda}$  is significantly different enough from zero to be included in the model.

So far, data series defined on a selected fixed set of sampling (time or space) points  $\tau_i$  have been dealt with. In some instances, one may want to regard one component of the series, usually the trend or the statistically dependent noise  $s$ , as being meaningful at other than just the sampling points  $\mathcal{T}$ . Usually, one wants to be able to speak of the series as being a discrete sample of functional values belonging to a function of a continuous parameter  $\tau$  of which  $\tau_i$ ,  $i=1,\dots,N$ , are only some selected values. The extension of the validity of such a series to points other than sampling points is called *prediction*. These other points are sometimes called *prediction points* and constitute a region  $\mathcal{P}$  which may or may not contain the sampling points as well, i.e.,  $\mathcal{T}$  may or may not be contained in  $\mathcal{P}$ . To predict a component of

the series, the component has to be regarded as a signal, i.e., as containing information useful enough that it should be modelled (cf. §10.3). The right-hand side of (8) can serve as an example of the prediction (predictor) of  $t$ , if the approximant is defined on domain  $\mathcal{P}$  such that  $\mathcal{T} \subset \mathcal{P}$ . Another, altogether different, example will be given in §14.3.

One more operation that can be performed on data series is *smoothing*, which is defined as the operation of separating signal from noise [GOLDMAN, 1953]. Since the definition of signal and noise depends on the interpretation of what is the useful part of the data series and what is not, then even the term ‘smoothing’ does not have a unique meaning. Consequently, any kind of approximation can be regarded as an operation of smoothing. For example, the approximant  $\Phi^T(\mathcal{T})\hat{\lambda}$  may be regarded as the smooth part of the analysed data series.

*Filtering* is a close relative of smoothing; it is actually defined as automatic smoothing [GOLDMAN, 1953] and the term is also being used loosely for smoothing. The general expression describing the filtering process is

$$\mathbf{l}' = \mathbf{f}(\mathbf{l}), \quad (14.15)$$

where  $\mathbf{l}'$  is the *smoothed data series*, and the function  $\mathbf{f}$  is known as the *filter*.

The most widely used filter is the *linear filter*, where the smoothed series is a linear function of the original series. It can be written as

$\mathbf{l}' = \mathbf{F}\mathbf{l},$

(14.16)

where the filter  $\mathbf{F}$  is simply a matrix. It can easily be understood that, again, the least-squares regression can be regarded as linear filtering, if one is willing to view the trend  $t$  as the smoothed version of series  $\mathbf{l}$ . Substituting for  $\hat{\lambda}$  from (12) into (8), we obtain

$$\mathbf{l}' = \mathbf{t} = \Phi^T(\mathcal{T})\tilde{\mathbf{G}}^{-1}\Phi(\mathcal{T})\mathbf{C}_r^{-1}\mathbf{l} = \mathbf{F}\mathbf{l}, \quad (14.17)$$

where  $\mathbf{F}$  depends only on  $\mathcal{T}$ ,  $\mathbf{C}_r$ , and on the choice of  $\phi$ .

Linear filters may be convolutive or recursive [GOLD AND RADER, 1969; OTNES AND ENOCHSON, 1972]. The *convolutive filter* works so that each value  $\mathbf{l}'(\tau_i)$  is given as a convolution of  $\mathbf{l}$ 's with another vector of values, i.e., as a linear combination of  $\mathbf{l}$ 's. In the case of (16), the linear combination is

$$\mathbf{l}'(\tau_i) = \sum_{j=1}^N f_{ij} \mathbf{l}(\tau_j). \quad (14.18)$$

*Recursive filters* are characterized by the following equation [GODIN, 1972]:

$$\mathbf{l}'(\tau_i) = \mathbf{F}_1(\mathbf{F}_2(\cdots(\mathbf{F}_k \mathbf{l}))), \quad (14.19)$$

which indicates a multiple application of linear filters  $\mathbf{F}$ , which may or may not be the same. Evidently, recursive filters can be regarded as a special kind of convolutive filters, where the convolutive filter is the product of recursive filters.

The main application of filters is in the domain of *oscillatory phenomena*, i.e., in the domain of repetitive or periodic characteristics. In many applications, the noise that is to be filtered out can be assumed to behave in an oscillatory fashion, being composed of one or several sinusoidal curves with various frequencies contained in a certain frequency band. In such a case, it is possible to design a convolutive filter of a sequential character; every value of the filtered series  $l'(\tau_i)$  is obtained as a specific linear combination of a part  $l(\tau_{i-k}), \dots, l(\tau_{i+m})$  of the original data series. To be able to design such a procedure, both the original and filtered data series must be equally spaced; i.e., the sampling points  $\tau_i$  may be expressed as  $\tau_i = \tau_0 + i\Delta\tau$ ,  $i=0, \dots, N$ , where  $\tau_0$  and  $\Delta\tau$  are constant. Then the *sequential filter* can be written as

$$l'(\tau_i) = \sum_{j=-k}^m f_{j+k+1} l(\tau_{i+j}), \quad i=0, \dots, N. \quad (14.20)$$

Note that elements  $l'(\tau_0), \dots, l'(\tau_{k-1}), l'(\tau_{N-m+1}), \dots, l'(\tau_N)$  of the filtered series are not defined, and thus we lose  $k$  elements at the beginning and  $m$  elements at the end. In matrix form, such a sequential filter can be shown as follows:

$$\begin{bmatrix} 0 \\ \vdots \\ k-1 \\ k \\ \vdots \\ N-m \\ N-m+1 \\ \vdots \\ N \end{bmatrix} \begin{bmatrix} 0 \\ I' \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ f_1, f_2, \dots, f_{k+m+1} \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} I \end{bmatrix}, \quad (14.21)$$

where the sequence of  $k+m+1$  elements  $f_j$  in this matrix remains the same for all the rows.

Two special kinds of sequential filters are used most widely in practice: predictive and symmetrical. *Predictive filters* are those which use only the past values of the original series  $I$  to obtain the filtered series  $I'$ , i.e.,  $m \leq 0$  in (20). These filters do not lose any elements at the end of the series. One example of such a filter for  $m=0$  is as follows:

$$\begin{bmatrix} 0 \\ \vdots \\ k \\ \vdots \\ I' \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} I \end{bmatrix}. \quad (14.22)$$

*Symmetrical filters*, on the other hand, are defined as having  $m = k$ ; and  $f_q = f_{-q}$  for  $q = 1, \dots, k$ . They shorten the filtered series by the same number of elements at both ends.

A linear sequential filter is said to be *normalized* if the condition

$$\sum_{j=-k}^m f_{i+k+j} = 1 \quad (14.23)$$

is satisfied. In the case of normalized filters, the amplitude of the signal  $l'$  obtained from the original series, is not distorted. This, indeed, is a very desirable property. Note that a normalized filter can be viewed as averaging the original series  $l$  within the interval  $\langle i - k, i + m \rangle$  using proportionate weights  $f_j$ , and so use of these filters is sometimes called the *technique of moving averages*. The performance of a simple normalized symmetrical filter, defined as

$$l'(\tau_i) = 0.1l(\tau_{i-2}) + 0.2l(\tau_{i-1}) + 0.4l(\tau_i) + 0.2l(\tau_{i+1}) + 0.1l(\tau_{i+2}), \quad (14.24)$$

is shown in FIG. 2.

It is often advantageous to regard the filter as a device, 'black box', into which the original series  $l$  is fed at one end and the smoothed signal  $l'$  appears at the other end. This is shown schematically in FIG. 3. In this context,  $l'$  is spoken of as the *response of the filter* to the original series  $l$  [GOLDMAN, 1953]. Often we speak of the *phase of the response* of the filter: the response can evidently be in phase with, in advance of, or lagging behind the original series. It is not difficult to understand that a symmetrical filter is the only one that introduces no phase distortion. On the other hand, a predictive filter introduces a phase lag: its response is phase distorted.

As stated earlier, since the main deployment of filters is in the domain of oscillatory phenomena, it is important to know which frequencies can be filtered out (erased) through the use of different types of filters. In other words, what does the response of a filter look like in the frequency domain? In order to answer this question and thus give some guidelines for designing filters with specific frequency characteristics, one must first learn something about spectral analysis.

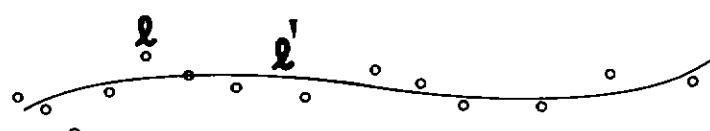


FIG. 14.2. Performance of a normalized symmetrical filter.

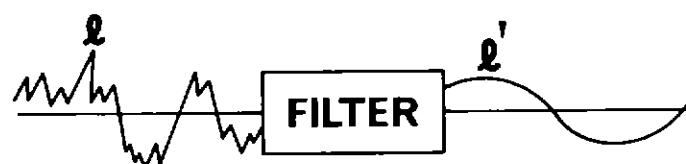


FIG. 14.3. Filter as a "black box".

Let us assume that the trend  $t$  of the series can be wholly or, at least, partly modelled by trigonometrical functions with various periods (frequencies) and amplitudes. In such a case,

$$t(\tau_j) = \sum_{s=1}^k (a_s \cos \omega_s \tau_j + b_s \sin \omega_s \tau_j) + \sum_{s=1}^m \lambda_s \phi_s(\tau_j). \quad (14.25)$$

Evidently, if the frequencies  $\omega_s$ ,  $s = 1, \dots, k$ , are known, the problem of determining  $t$  is reduced to the determination of the vector of coefficients  $\lambda' = [a_1, b_1, \dots, a_k, b_k; \lambda_1, \dots, \lambda_m]^T$ , where  $\dim(\lambda') = u = 2k + m$ . Thus, denoting  $[\cos \omega_1 \tau, \sin \omega_1 \tau, \dots, \cos \omega_k \tau, \sin \omega_k \tau; \phi_1(\tau), \dots, \phi_m(\tau)]$  by  $\Phi'^T(\tau)$ , one has, from (8),

$$t(\tau) = \tilde{\Phi}'^T(\tau) \lambda', \quad (14.26)$$

which is, again, the problem of linear regression already treated.

The situation becomes much more complicated when the frequencies in (25) are not known beforehand. Since the frequencies in (25) are built into the trigonometrical functions, we have a non-linear problem on our hands and, therefore, a solution does not generally exist (cf. §11.1). In order to even attempt a solution, good first approximations of the frequencies, known or suspected to be present in the signal, must be available. This is normally not the case, and the task of determining these unknown frequencies is usually solved by means of *spectral analysis*, which is a term covering a whole family of techniques.

The basic technique used for an approximate spectral analysis is *harmonic analysis* (e.g., ZYGMUND [1968]). To understand how harmonic analysis is used in this context, let us assume there is a data series  $l(\tau_j)$ , observed with equal accuracies and composed only of a trend (signal), that can be expressed as a *trigonometrical polynomial*, i.e.,

$$l(\tau_j) = a_0 + \sum_{s=1}^k (a_s \cos s \tau_j + b_s \sin s \tau_j), \quad (14.27)$$

where  $a_0$  is the constant term. In addition, suppose that the observations  $l$  were sampled on  $2n$  equidistant points; namely,

$$\mathcal{T} \equiv \{-\pi + (\pi/n)j; j = 1, \dots, 2n\}. \quad (14.28)$$

This is not a serious restriction since any argument consisting of  $2n$  equidistant points, say  $\{\tau'_j = c + ((d - c)/2n)j; j = 1, \dots, 2n\}$ , can be reduced to the above through a simple linear transformation:

$$\tau_j = \frac{2\pi}{d - c} (\tau'_j - c) - \pi, \quad j = 1, \dots, 2n. \quad (14.29)$$

It has been shown by, e.g., LANCZOS [1957] that on this particular  $\mathcal{T}$  the set of functions,

$$\begin{aligned} \tilde{\phi}(\tau) &\equiv \{1, \cos s \tau, \sin s \tau; s = 1, \dots, k < \frac{1}{2}(n-1)\} \\ &\equiv \{\phi_0(\tau), \phi_1(\tau), \dots, \phi_{2k}(\tau)\}, \end{aligned} \quad (14.30)$$

is an orthogonal base for the metric equal to  $I$ . Hence, the best-fitting coefficients in the least-squares sense,

$$\hat{\lambda} \equiv \{\hat{a}_0, \hat{a}_s, \hat{b}_s; s = 1, \dots, k\}, \quad (14.31)$$

can be determined from the normal eqn. (12). Further, on the strength of the orthogonality of  $\phi$ , the normal equations reduce to eqn. (14). Evidently we have

$$u_s = - \sum_{j=1}^{2n} l(\tau_j) \phi_s(\tau_j), \quad s = 0, \dots, 2k, \quad (14.32)$$

and it can be proved that

$$g_{ss} = \begin{cases} 2n, & s = 0, \\ n, & s > 0. \end{cases} \quad (14.33)$$

Thus one gets

$$\begin{aligned} \hat{a}_0 &= \frac{1}{2n} \sum_{j=1}^{2n} l(\tau_j) = \frac{1}{2n} \sum_{j=1}^{2n} l(\tau_j) \cos s \tau_j, & s = 0, \\ \hat{a}_s &= \frac{1}{n} \sum_{j=1}^{2n} l(\tau_j) \cos s \tau_j & (14.34) \\ && s = 1, \dots, k. \\ \hat{b}_s &= \frac{1}{n} \sum_{j=1}^{2n} l(\tau_j) \sin s \tau_j \end{aligned}$$

From trigonometry it is known that the following relation can be written for the  $s$ th trigonometrical term in (27):

$$a_s \cos s \tau + b_s \sin s \tau = A_s \cos(s \tau - \psi_s), \quad (14.35)$$

where  $A_s$  is the *amplitude* of the term and  $\psi_s$  is its *phase*. This is why the trigonometrical term is often called a *wave*. The relationship between the coefficients on the one hand and the amplitude and phase on the other can be derived from (35) and the tangent of half angle formula as

$$A_s = \sqrt{a_s^2 + b_s^2}, \quad \psi_s = 2 \arctan [b_s / (A_s + a_s)]. \quad (14.36)$$

It is illustrative to plot the derived amplitudes against the integral frequencies  $s$  (see FIG. 4) of the individual waves. Naturally, the larger the amplitude, the greater the wave's contribution to the observed series. Thus the plot can serve as an indicator of the importance of individual waves. In some applications, a similar plot of  $\psi$  against  $s$  is used. Note that the amplitude  $A$  is plotted as a function of frequency  $s$ ; thus it can also be regarded as a transformation of the original series  $l(\tau)$  into a new space spanned by the frequencies—the *frequency space*. The explicit expression for the transformation from the space spanned by the argument  $\tau$  (i.e., geometrical or time space) into the frequency space is obtained by combining (36)

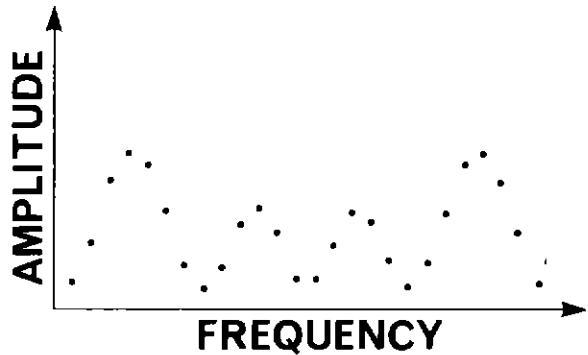


FIG. 14.4. Harmonic analysis.

and (34):

$$\hat{A}(s)|_t = \frac{1}{n} \sqrt{\left( \sum_{j=1}^{2n} l(\tau_j) \cos s\tau_j \right)^2 + \left( \sum_{j=1}^{2n} l(\tau_j) \sin s\tau_j \right)^2}, \quad s = 0, \dots, k. \quad (14.37)$$

The result of such a transformation, in the frequency space, is known as the *Fourier spectrum of  $l$* .

In nature, the various phenomena affecting the observable are unlikely to have integral frequencies over any selected sampling interval  $\langle c, d \rangle$ . If the frequencies are not integral, the harmonic analysis gives distorted results because the model, given by (27), does not describe the series adequately. To overcome this problem, it is natural to expand the idea of harmonic analysis to non-integral frequencies  $\omega_s$ . The trigonometrical polynomial in the model given by (27) is replaced by the generalized trigonometrical polynomial, such as the one we saw in (25). The corresponding transformation equation into the compact frequency space then becomes, by analogy with (37),

$$A(\omega)|_t = \frac{1}{n} \sqrt{\left( \sum_{j=1}^{2n} l(\tau_j) \cos \omega \tau_j \right)^2 + \left( \sum_{j=1}^{2n} l(\tau_j) \sin \omega \tau_j \right)^2}. \quad (14.38)$$

This equation is known as the discrete *Fourier transformation*. It is continuous for all values of  $\omega \in \langle 0, n \rangle$ . The highest frequency for which the transformation is normally sought is  $\omega = n$ , called the *folding, or Nyquist, frequency*. The spectrum obtained by means of the Fourier transformation is called the *Fourier spectrum, or the periodogramme*. An example of such a periodogramme, corresponding to the results of harmonic analysis shown in FIG. 4, is seen in FIG. 5; another close relative of a periodogramme was shown in FIG. 8.30.

In many applications, it is convenient to express the Fourier transformation in complex form. This is normally done through the following definition:

$$Z^*(\omega) = A(\omega) \exp[-i\psi(\omega)] = (a(\omega), -b(\omega)), \quad (14.39)$$

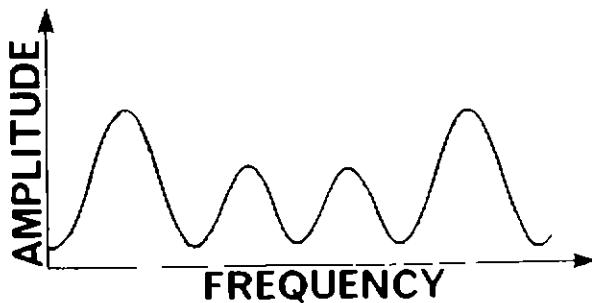


FIG. 14.5. Periodogramme.

where  $i = \sqrt{-1}$  and  $a(\omega), b(\omega)$  are given as (cf. (34))

$$a(\omega) = \frac{1}{n} \sum_{j=1}^{2n} l(\tau_j) \cos \omega \tau_j, \quad b(\omega) = \frac{1}{n} \sum_{j=1}^{2n} l(\tau_j) \sin \omega \tau_j. \quad (14.40)$$

It can be seen immediately that the spectrum is the absolute value of the complex function  $Z^*(\omega)$  (see §3.1); i.e.,

$$A(\omega) = \sqrt{a^2(\omega) + b^2(\omega)}, \quad (14.41)$$

and the phase is the argument of  $Z^*(\omega)$ ; i.e.,

$$\psi(\omega) = 2 \arctan \{ b(\omega) / [A(\omega) + a(\omega)] \}. \quad (14.42)$$

Expressing  $\cos \omega \tau$  and  $\sin \omega \tau$  now in complex form, using Moivre's theorem (see §3.1), one obtains the *complex Fourier transformation* as

$$Z^*(\omega) = \frac{1}{n} \sum_{j=1}^{2n} l(\tau_j) \exp(-i\omega\tau_j). \quad (14.43)$$

The complex conjugate of the complex Fourier transformation reads:

$$\bar{Z}^*(\omega) = \frac{1}{n} \sum_{j=1}^{2n} l(\tau_j) \exp(i\omega\tau_j). \quad (14.44)$$

Using the complex conjugate, the following relations can also be written:

$$A(\omega) = \sqrt{Z^*(\omega) \bar{Z}^*(\omega)}, \quad (14.45)$$

$$\psi(\omega) = \frac{1}{2i} \ln [Z^*(\omega) / \bar{Z}^*(\omega)].$$

One would like to think that the peaks of a periodogramme indicate the frequencies (now non-integral) which contribute the most to the analysed data series; i.e., the frequencies sought by means of the spectral analysis. Strictly speaking, this would be the case only for an infinitely long data series. The finiteness of the actually observed series always causes a certain distortion of the spectrum. The

finiteness also causes the broadening of spectral peaks that would otherwise be infinitely thin; the spectrum of an infinitely long series composed of only trigonometrical terms is thus a *line spectrum*, an example of which was seen in FIG. 8.4. Other sources of distortions are the interference (beat) and aliasing of different frequencies. For a discussion and explanation of these phenomena, the reader is referred to the literature (e.g., BLACKMAN AND TUKEY [1958], JENKINS AND WATTS [1968]).

There is one class of distortions, however, that should be discussed here: the distortions caused by the presence of other than trigonometrical functions in the complete model of the data series. Suppose that the series can be decomposed into three parts as follows:

$$l = \hat{l} + p - r, \quad (14.46)$$

where  $\hat{l}$  is the pure, generalized, trigonometrical polynomial;  $p$  is another polynomial, such as the second term in (25); and  $r$  is the residual. From the point of view of spectral analysis, both  $p$  and  $r$  may be regarded as noise. Since the complex Fourier transformation of  $l$  is a linear function of  $l$  (cf. (43)), we have (cf. §3.1)

$$Z^*(\omega)|_l = Z^*(\omega)|_{\hat{l}} + Z^*(\omega)|_p - Z^*(\omega)|_r. \quad (14.47)$$

Consequently, the spectrum  $A(\omega)|_l$  will reflect the effect of both noise components  $p$  and  $r$ . Looking at the effect of  $r$  first, the spectrum of the residual  $r$  will depend on the properties of  $r$ . If the *random noise*  $r$  is statistically independent, it will have, by definition, a constant or *flat spectrum*; such an  $r$  is referred to as a *white noise*. In some textbooks, the flatness of the spectrum is what distinguishes a purely random from a non-random series. If  $r$  is not purely random, then it has to be treated as non-white, or systematic, noise and brought together with  $p$ . Either way, it will be disregarded in the ensuing discussion.

The component  $p$  may be viewed as *systematic noise*. Its complex Fourier transformation can be written as

$$Z^*(\omega)|_p = \sum_{s=1}^m \lambda_s Z^*(\omega)|_{\phi_s}, \quad (14.48)$$

because the transformation is linear, as shown above. In order to correct for its effect, the magnitudes  $\lambda_s$  and the transforms  $Z^*(\omega)|_{\phi_s}$  of the individual functions have to be known. Although transforms for different shapes of functions  $\phi_s$  [GODIN, 1972] are relatively easy to obtain, the magnitudes are not normally known. If they were, then the systematic noise could be subtracted from the data series  $l$ , and its effect would be removed. Usually, the magnitudes  $\lambda$  are estimated beforehand, and the estimated  $p$  is subtracted from  $l$ . This is rather unsatisfactory because the accuracy of the estimated  $\lambda$  still leaves the spectrum of the corrected  $l$  contaminated. A more satisfactory solution is obtained using *least-squares spectral analysis* [VANÍČEK, 1971], which eliminates the ill-effect of the ignorance of magnitudes. An example of the least-squares spectrum was shown in FIG. 8.33.

Let us now turn to the relation between linear filters and Fourier spectral analysis. It should be obvious by now that any spectral analysis is helpful in the first

determination of the limits (band) of frequencies, present in the series, which one may want to filter out. In this context, of course, the periodic part of the series that is to be filtered out is to be considered as noise. More importantly, the response of a filter can be assessed in the frequency space using the Fourier transformation. It can be shown (e.g., JENKINS AND WATTS [1968]) that a complex Fourier transformation of a sequentially filtered series (cf. (20)) is given as

$$Z^*(\omega)|_r = Z^*(\omega)|_t \cdot Z^*(\omega)|_f, \quad (14.49)$$

where  $Z^*(\omega)|_f$  is the Fourier transformation of the coefficients of the filter which is, in this instance, viewed as a numerical function defined on any subset of adjacent sampling points from the set  $\mathfrak{T}$ . The effect of a filter on the spectrum of the filtered series is thus given, realizing that the amplitude of a product of two complex numbers is the product of amplitudes of these two numbers (cf. §3.1), by the following equation:

$$A(\omega)|_r = A(\omega)_t \cdot A(\omega)|_f. \quad (14.50)$$

Any selected sequence of coefficients  $f_i$  has a spectrum  $A(\omega)|_f$  that determines the response of the linear filter represented by these coefficients. As an example, let us again take the simple filter given by (24). The coefficients  $f_{-2} = f_2 = 0.1$ ,  $f_{-1} = f_1 = 0.2$ ,  $f_0 = 0.4$  interpreted as a numerical function are shown in FIG. 6. Note that the location of the five non-zero values on the  $\tau$  axis is irrelevant as far as the spectrum is concerned. The spectrum of this function is plotted in FIG. 7. It can be seen that this particular filter does not affect the low-frequency waves, but attenuates or eliminates the high-frequency waves. Filters with similar characteristics are called *low-pass filters*. In practice, *high-pass* and *band-pass filters* are also used.

In conclusion, let us mention the fact that the Fourier transformation is not the only available technique of spectral analysis. In addition to the already pointed out least-squares spectral analysis there is a variety of techniques having special features convenient for analysing series with particular properties. Although any discussion of these would be out of the question here, the reader should be made aware of at

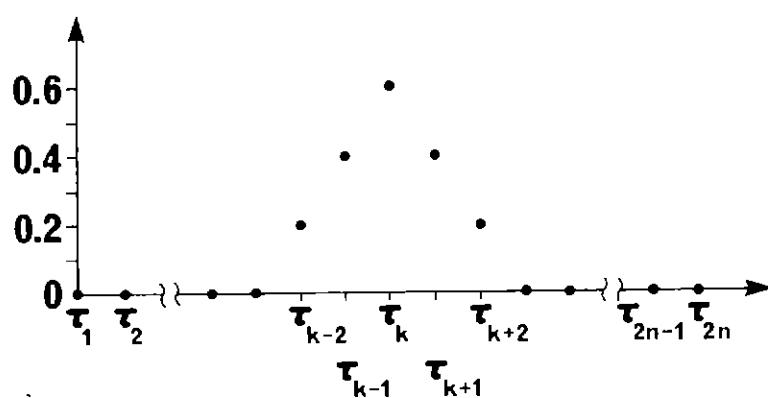


FIG. 14.6 Linear, normalized, symmetrical filter as a numerical function.

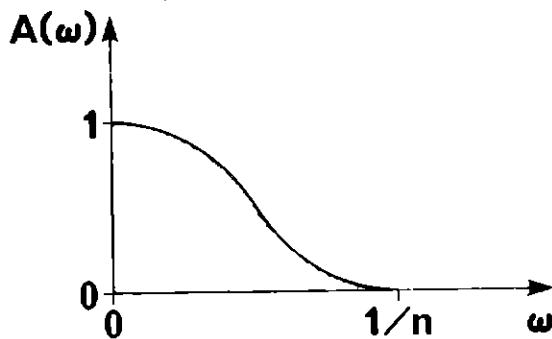


FIG. 14.7. Spectrum of the low-pass filter shown in Fig. 6.

least the *power spectrum technique* [JENKINS AND WATTS, 1968], the *maximum cutoff* [BURG, 1967], and the *maximum likelihood* [CAPON, 1969] methods. Techniques also exist for simultaneous investigation of two or several data series. These are treated in standard textbooks; e.g., WILKS [1962], JENKINS AND WATTS [1968].

### 14.3. Adjustment of observations

The *simple adjustment* of statistically independent observations  $\mathbf{l}$  uses the following mathematical model:

$$\boxed{f(\mathbf{x}, \mathbf{l} + \mathbf{v}) = \mathbf{0}}, \quad (14.51)$$

and is thus characterized by the absence of  $\mathbf{s}$ . The simple adjustment consists of the least-squares estimation of the unknown parameters and the correction (adjustment) of the observations to make them consistent within the framework of the model. Thus, after the following equivalences are assumed,  $\mathbf{C}_l \equiv \mathbf{C}_r \equiv \mathbf{C}_v \equiv \text{diag}(\sigma_i^2)$ , all the equations in Chapter 12 can be applied immediately. The *adjustment* of statistically dependent observations is characterized through the metricalization of  $\mathcal{L}$  by means of the inverse of a fully populated covariance matrix of the observations. This implies  $\mathbf{C}_l \equiv \mathbf{C}_r \equiv \mathbf{C}_s$ , recalling that, by definition,  $\mathbf{C}_s$  is fully populated (cf. §10.4). The situation of having both random components,  $\mathbf{v}$  and  $\mathbf{s}$ , present can be treated either by keeping them separated, as will be shown a little later, or by combining them into one. In the latter case, the covariance matrix corresponding to the combined residual  $\mathbf{r}$  is the sum of the two corresponding covariance matrices (cf. (10.44)), and, once again, one is in the realm of adjustment.

A complication arises if the systematic effects on the observations have not been successfully removed at the preprocessing stage. It then becomes necessary to remove the remaining effects at the time the observations are being adjusted to fit into the main mathematical model. This is accomplished simply by introducing into the model some unknown nuisance parameters  $\lambda$  in the form of an additional term  $\Phi^T(\mathcal{T})\lambda$  (cf. (8)). The end result is a *simultaneous adjustment and regression* with the

following least-squares normal equations:

$$\begin{bmatrix} \mathbf{N}_x & | & \mathbf{N}_{x\lambda} \\ \hline \mathbf{N}_{\lambda x} & | & \mathbf{N}_{\lambda} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\boldsymbol{\lambda}} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_x \\ \mathbf{u}_{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad (14.52)$$

the formulation and solution of which is discussed in detail in a slightly different context in §14.4.

Turning to the problem of *two-component adjustment* of observations, let us consider, without any loss of generality, the implicit mathematical model the general form of which is (cf. (12.1))

$$\boxed{\mathbf{f}(\mathbf{x}, \hat{\mathbf{l}}) = \mathbf{f}(\mathbf{x}, \mathbf{l} + \mathbf{s} + \mathbf{v}) = \mathbf{0}}, \quad (14.53)$$

and the covariance matrices  $\mathbf{C}_s$  and  $\mathbf{C}_v$ . The cross-covariance matrix  $\mathbf{C}_{sv}$  will be assumed to be equal to zero. The differential form of the model (cf. (12.2)) can be written as

$$\mathbf{A}\boldsymbol{\delta} + \mathbf{B}\mathbf{s} + \mathbf{B}\mathbf{v} + \mathbf{w} = \mathbf{0}, \quad (14.54)$$

if the superscript (0) is dropped from  $\mathbf{w}$ . In the more general case, the component  $\mathbf{s}$  is not from the observation space  $\mathcal{L}$  but from another space of statistically dependent observations  $\mathcal{S}$ , such that for the differential neighbourhood of the expansion point  $\mathbf{s}$ , the following transformation holds (see FIG. 8):

$$\mathbf{T} \in \{\mathcal{S} \rightarrow \mathcal{L}\}. \quad (14.55)$$

Let us denote the original  $\mathbf{s}$  by  $\mathbf{s}'$ , then we have

$$\mathbf{s}' = \mathbf{T}\mathbf{s}, \quad (14.56)$$

where  $\mathcal{S}$  may or may not have the same dimension as  $\mathcal{L}$ . By substituting (56) into (54), and denoting  $\mathbf{B}$  by  $\mathbf{B}_v$  and  $\mathbf{BT}$  by  $\mathbf{B}_s$ , one gets the more general, two-component linear model:

$$\boxed{\mathbf{A}\boldsymbol{\delta} + \mathbf{B}_v\mathbf{v} + \mathbf{B}_s\mathbf{s} + \mathbf{w} = \mathbf{0}}. \quad (14.57)$$

The variation function from which the normal equation system follows, reads

$$\phi = \mathbf{s}^T \mathbf{C}_s^{-1} \mathbf{s} + \mathbf{v}^T \mathbf{C}_v^{-1} \mathbf{v} + 2\mathbf{k}^T (\mathbf{A}\boldsymbol{\delta} + \mathbf{B}_s\mathbf{s} + \mathbf{B}_v\mathbf{v} + \mathbf{w}). \quad (14.58)$$

Clearly, both sets of residuals play their roles in the quadratic forms, along with their respective covariance matrices; it is illustrative to compare this equation with (12.12). The normal equations are obtained along the lines followed in §12.2. The quantities  $\boldsymbol{\delta}$ ,  $\hat{\mathbf{x}}$ ,  $\mathbf{C}_{\hat{\mathbf{x}}}$ , and  $\hat{\mathbf{k}}$  are then evaluated from (12.26), (12.11), (12.36), and (12.28), where  $\mathbf{C}_r$  and  $\mathbf{B}$  are replaced by  $\mathbf{C}'_r$ :

$$\mathbf{C}'_r = \begin{bmatrix} \mathbf{C}_s^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{C}_v^{-1} \end{bmatrix}, \quad (14.59)$$

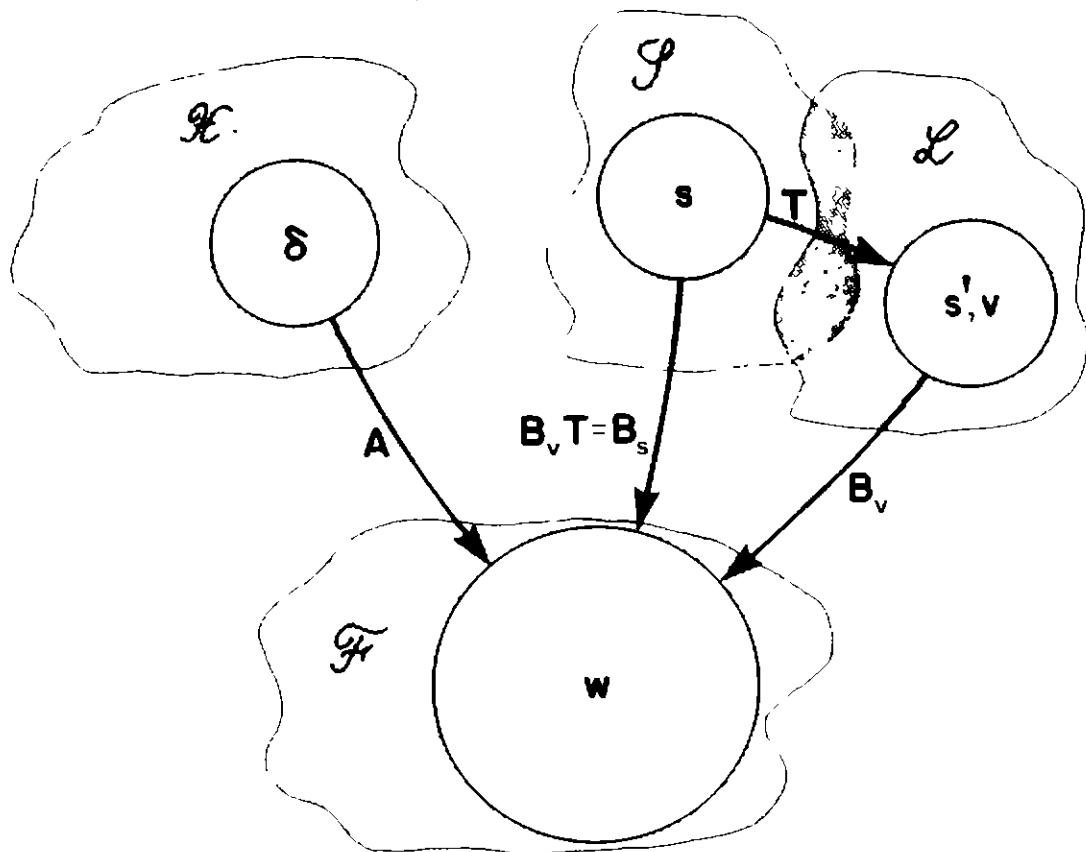


FIG. 14.8. Spaces used in two-component adjustment of observations.

and  $\mathbf{B}'$ :

$$\mathbf{B}' = [\mathbf{B}_s \mid \mathbf{B}_v]. \quad (14.60)$$

Realizing that  $\hat{\mathbf{r}}' = [\hat{s} \mid \hat{v}]^T$  one obtains, from (12.29),

$$\hat{\mathbf{r}}' = - \begin{bmatrix} \mathbf{C}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_v \end{bmatrix} \begin{bmatrix} \mathbf{B}_s^T \\ \mathbf{B}_v^T \end{bmatrix} \hat{\mathbf{k}}, \quad (14.61)$$

and

$$\hat{s} = -\mathbf{C}_s \mathbf{B}_s^T \mathbf{L} \mathbf{w}, \quad (14.62)$$

$$\hat{v} = -\mathbf{C}_v \mathbf{B}_v^T \mathbf{L} \mathbf{w}. \quad (14.63)$$

Their covariance matrices, again, follow immediately from (12.38) as

$$\mathbf{C}_{\hat{\mathbf{r}}}' = \begin{bmatrix} \mathbf{C}_{\hat{s}} & \mathbf{C}_{\hat{s}\hat{v}} \\ \mathbf{C}_{\hat{v}\hat{s}} & \mathbf{C}_{\hat{v}} \end{bmatrix} = \mathbf{C}_r \mathbf{B}'^T \mathbf{L} \mathbf{B}' \mathbf{C}_r'.$$

Specifically,

$$\mathbf{C}_{\hat{s}} = \mathbf{C}_s \mathbf{B}_s^T \mathbf{L} \mathbf{B}_s \mathbf{C}_s, \quad (14.64)$$

$$\mathbf{C}_{\hat{v}} = \mathbf{C}_v \mathbf{B}_v^T \mathbf{L} \mathbf{B}_v \mathbf{C}_v, \quad (14.65)$$

$$\mathbf{C}_{\hat{s}\hat{v}} = \mathbf{C}_{\hat{v}\hat{s}} = \mathbf{C}_s \mathbf{B}_s^T \mathbf{L} \mathbf{B}_v \mathbf{C}_v = (\mathbf{C}_v \mathbf{B}_v^T \mathbf{L} \mathbf{B}_s \mathbf{C}_s)^T. \quad (14.66)$$

In all these equations, the effect of the addition of the second residual is felt through the matrix  $\mathbf{L}$ . Equations (59) and (60) directly affect  $\mathbf{M}$ , changing it to

$$\mathbf{M} = \left( \begin{bmatrix} \mathbf{B}_s & \mathbf{B}_v \end{bmatrix} \begin{bmatrix} \mathbf{C}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_v \end{bmatrix} \begin{bmatrix} \mathbf{B}_s^T \\ \mathbf{B}_v^T \end{bmatrix} \right)^{-1} = (\mathbf{B}_s \mathbf{C}_s \mathbf{B}_s^T + \mathbf{B}_v \mathbf{C}_v \mathbf{B}_v^T)^{-1}, \quad (14.67)$$

and through  $\mathbf{M}$  the changes get transformed to  $\mathbf{N}$  and  $\mathbf{L}$ . Finally, it is noted that the a posteriori variance factor  $\hat{\sigma}_0^2$  is, again, evaluated from (12.51), when  $\hat{\mathbf{r}}$  and  $\mathbf{P}_t = \sigma_0^2 \mathbf{C}_r^{-1}$  are, of course, changed to  $\hat{\mathbf{r}}'$  and  $\mathbf{P}'_t = \sigma_0^2 \mathbf{C}'_r^{-1}$ . A situation may even arise where it would be desirable to treat several random variables separately. Such a case will not be dealt with here: it would simply be a direct generalization of the problem treated above. It is of interest to look at the results of the two-component adjustment from the point of view of data series analysis. The sum

$$\mathbf{I}' = \mathbf{I} + \hat{\mathbf{v}} \quad (14.68)$$

may be called the smoothed observation series. Substituting for  $\hat{\mathbf{v}}$  from (63) and, again, for  $\mathbf{w}$  from (12.3), one gets

$$\mathbf{I}' = \mathbf{I} - \mathbf{C}_v \mathbf{B}_v^T \mathbf{L} \mathbf{f}(\mathbf{x}^{(0)}, \mathbf{I}), \quad (14.69)$$

which represents the equation of a special filter—to be called a *covariance filter*.

The two-component adjustment combined with a prediction of the signal  $s$  is, in the literature, called *least-squares collocation* [KRARUP, 1969]. From the point of view of smoothing and filtering, the two-component adjustment is the same as the two more simple kinds of adjustments treated earlier in this section. Either of these kinds of adjustment separates the signal ( $\mathbf{x}$ ) from the noise ( $\mathbf{v}$ , or  $s$ , or  $\mathbf{v} + s$ ). What distinguishes the least-squares collocation is that it views the statistically dependent component  $s$  also as a signal, and one thus has two kinds of signal,  $\mathbf{x}$  and  $s$ . The first signal ( $\mathbf{x}$ ) does not naturally lend itself to prediction because  $\mathbf{x}$  is defined on a parameter space  $\mathcal{X}$  usually reflecting some specific observables, and thus it does not make any physical sense to predict  $\mathbf{x}$  elsewhere. On the other hand,  $s$  is often thought of as being only a sample of an effect that can be modelled in a wider space  $\mathcal{P}$  than just the observation space  $\mathcal{S}$ . Thus, a prediction, denoted by  $s_p$ , of the statistically dependent signal  $s$  can be sought in a *prediction space*  $\mathcal{P}$  such that  $\mathcal{S} \subset \mathcal{P}$  (see FIG. 9).

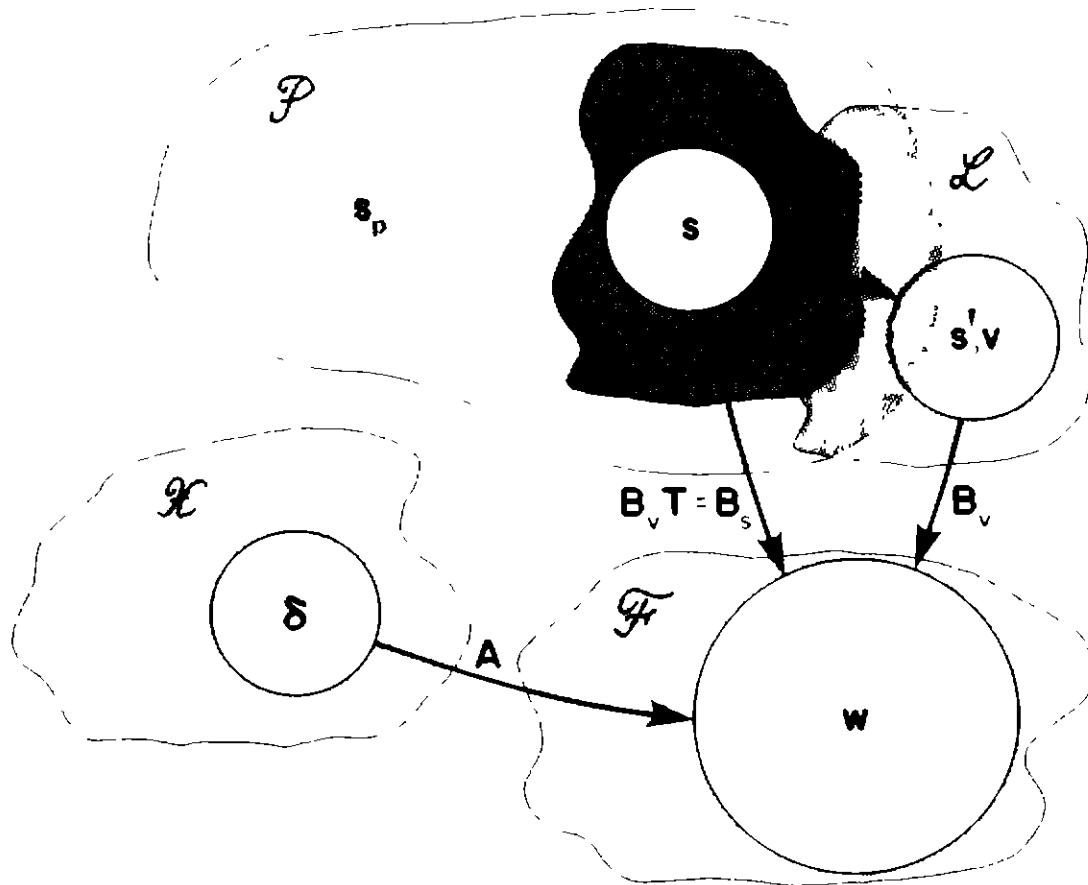


FIG. 14.9. Spaces used for prediction by least-squares collocation.

Mathematically, the prediction feature can be built into the two-component adjustment by simply expanding the second design matrix to  $\mathbf{B}''$  as

$$\mathbf{B}'' = [\mathbf{0} \mid \mathbf{B}_s \mid \mathbf{B}_v], \quad (14.70)$$

and by taking

$$\hat{\mathbf{r}}'' = [\hat{s}_p^T \mid \hat{s}^T \mid \hat{v}^T]^T. \quad (14.71)$$

Note that the null matrix in the hyper-matrix  $\mathbf{B}''$  serves as the means by which the prediction is built into the scheme. The model (57) is unaffected by this mathematical trick; this can be verified by substituting the expressions for  $\mathbf{B}''$  and  $\hat{\mathbf{r}}''$  into (12.2) and making a comparison with (57).

The prediction is really made possible by the stipulation that the stochastical characteristics of the new quantity  $s_p$  are the same as those of  $s$ . This is expressed through an expanded covariance matrix  $\mathbf{C}'_s$

$$\mathbf{C}'_s = \begin{bmatrix} \mathbf{C}_{s_p} & \mathbf{C}_{s_p s} \\ \mathbf{C}_{s s_p} & \mathbf{C}_s \end{bmatrix}, \quad (14.72)$$

which describes the covariance between  $s$  at the observation (sample) points and  $s_p$  at the prediction points. Here,  $\mathbf{C}_{s_p s}$  equals  $\mathbf{C}_{s s_p}^T$  and thus preserves the required symmetry of  $\mathbf{C}'_s$ . It is the covariance among all the signal components, expressed in

terms of the covariance function (cf. (10.30)), that mediates the prediction by relating  $s$  to  $s_p$ .

All the equations for the two-component adjustment remain valid when the expanded versions of  $\mathbf{B}$  (70) and  $\mathbf{r}$  (71) are used. Changes only occur to (62) which, upon substitution of  $\mathbf{C}'_s$  for  $\mathbf{C}_s$ , both reproduces itself and yields a new equation—that of the predicted signal:

$$\hat{s}_p = -\mathbf{C}_{s_p s} \mathbf{B}_s^T \mathbf{L} \mathbf{w}. \quad (14.73)$$

Similarly, the same substitution into (64) not only reproduces the same equation but also gives two new equations: namely, that for the covariance matrix of the predicted signal,

$$\mathbf{C}_{\hat{s}_p \hat{s}_p} = \mathbf{C}_{s_p s} \mathbf{B}_s^T \mathbf{L} \mathbf{B}_s \mathbf{C}_{s s_p}, \quad (14.74)$$

and that for the cross-covariance matrix between the predicted signal and the estimated signal at the sampling points,

$$\mathbf{C}_{\hat{s}_p \hat{s}_v} = \mathbf{C}_{\hat{s}_p \hat{s}}^T = \mathbf{C}_s \mathbf{B}_s^T \mathbf{L} \mathbf{B}_s \mathbf{C}_{s s_p}. \quad (14.75)$$

By analogy with (66), the cross-covariance matrix between the predicted signal and the estimated, statistically independent residuals on sampling points can also be found to be

$$\mathbf{C}_{\hat{s}_p \hat{\epsilon}} = \mathbf{C}_{\hat{s}_p \hat{s}}^T = \mathbf{C}_v \mathbf{B}_v^T \mathbf{L} \mathbf{B}_s \mathbf{C}_{s s_p} = (\mathbf{C}_{s_p s} \mathbf{B}_s^T \mathbf{L} \mathbf{B}_v \mathbf{C}_v)^T. \quad (14.76)$$

Because the prediction feature does not affect the mathematical model, the unknown parameters  $x$  can be first estimated from adjustment with  $\mathbf{C}_r = \mathbf{C}_s + \mathbf{C}_v$ . Then  $s$  can be predicted afterwards using (73) based on the knowledge of the cross-covariance matrix  $\mathbf{C}_{s_p s}$ . It is interesting to note that even though  $\mathbf{C}_{s_p}$  is needed in the definition of  $\mathbf{C}'_s$  (72), i.e., it must be assumed to exist, it is not present in the formula for the predicted signal  $s_p$ . The reason is that  $\mathbf{C}_{s_p}$  does not contain any information useful from the prediction point of view. The a posteriori variance factor  $\hat{\sigma}_0^2$  in the case of least-squares collocation prediction is given by the usual expression (12.51). The fact that the prediction is made together with or after the adjustment, of course, makes no difference to the accuracy of the adjustment itself.

All of the expressions developed in this section are valid for an implicit model. If a model explicit in  $t$  is used instead, the only difference will be that  $\mathbf{B}_s = \mathbf{B}_v = -\mathbf{I}$ , and all the formulae will be considerably simplified. The matrix  $\mathbf{L}$ , for instance, becomes

$$\mathbf{L} = \mathbf{C}_r^{-1} - \mathbf{C}_r^{-1} \mathbf{A} (\mathbf{A}^T \mathbf{C}_r^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C}_r^{-1}. \quad (14.77)$$

If a condition model is used, the formulae are further simplified because  $\mathbf{A} = \mathbf{0}$  and

$$\mathbf{L} = \mathbf{C}_r^{-1}. \quad (14.78)$$

In both cases,  $\mathbf{C}_r = \mathbf{C}_t$  is given by (10.44).

To finish with, let us apply the ideas of this section to yet another problem, that of the *decomposition of a random series* composed of both statistically dependent and independent components with no cross-covariance between them. This is essentially a problem of separating the random signal (useful part) from the random noise and could be visualized by means of FIG. 9, where the solution space is left out. We have (cf. (10.36)):

$$l(\tau) = -r(\tau) = -s(\tau) - v(\tau). \quad (14.79)$$

In the vectorial form, the model is

$$l + s + v = \mathbf{0}. \quad (14.80)$$

Thus, using the terminology employed in this section,  $B_s = B_v = -I$ ,  $A = \mathbf{0}$ ,  $w = -l$ , and the model becomes clearly a special case of a condition model (cf. eqn. (10.7)). Equation (62) then gives the following expression for the estimated signal:

$$\hat{s} = -C_s C_r^{-1} l. \quad (14.81)$$

One application of the above equation would be to split into two parts a residual vector  $\hat{r}$  obtained from an adjustment where  $C_r = C_s + C_v$  had been used. This is another proof that the adjustment with  $C_r$  and the two-component adjustment (that uses  $C'$ ) can be done sequentially. Note that  $-\hat{s}$  ( $= l + \hat{v}$ ) is the same as the  $l'$  series smoothed by the covariance filter (69). The proof is left to the reader. The smoothed series then can be predicted in the  $\mathcal{P}$  space by

$$\hat{s}_p = -C_{s_p s} C_r^{-1} l. \quad (14.82)$$

This is known as the *Wiener–Kolmogorov* formula [LIEBELT, 1967] which represents a particularly simple and popular application of least-squares collocation.

Statistical testing of hypotheses involving variance factors  $\hat{\sigma}_0^2$ , components  $\hat{s}$ ,  $\hat{s}_p$ ,  $\hat{v}$ , and their covariance matrices can be performed using the methodology given in Chapter 13, whether these quantities originate from a simple adjustment, an adjustment, a two-component adjustment, or even an adjustment combined with the least-squares collocation prediction. The alteration of the appropriate formulae is simple and is, once more, left to the reader.

#### 14.4. Problems with a priori knowledge about the parameters

The adjustment techniques shown in §14.3 may be understood as tacitly assuming that the a priori value of the weight matrix  $C_x^{-1}$  equals to  $\mathbf{0}$ . The presence of this built-in assumption will be made clear in this section. The premise for this assumption is somewhat incorrect from the statistical point of view; one always knows, to a certain extent, what the value of  $x$  should approximately be. One has to know this value for linearization, where  $x^{(0)}$  is needed in the evaluation of all the design

matrices. Even if the model is linear, an approximate value of  $x$  may be obtained from the model when a minimal set of observations, needed to get one possible solution  $x$ , is used. Translation of this finding into the language of statistics tells us that a certain weight should be placed on  $x^{(0)}$  and thus  $C_x^{-1}$  should, from the beginning, be different from  $\mathbf{0}$ . The objective of this section is to show how to incorporate such a  $C_x^{-1} \neq \mathbf{0}$  into the adjustment, or regression, to get a more proper solution. Clearly, the spaces used are the  $\mathcal{L}$ ,  $\mathcal{X}$ , and  $\mathcal{F}$  of Chapter 12. The only difference is that  $\mathcal{X}$  is metricized right from the outset by  $C_x^{-1}$  rather than by a metric induced from  $\mathcal{L}$ .

As two distinctly different situations may occur, there are two different schools of thought on the subject: the Bayesian school [BAYES, 1763] that deals with the situation when  $C_x$  is subjectively chosen, and the generalized adjustment school [SCHMID AND SCHMID, 1965] that treats the situation when  $C_x$  is objectively known, i.e., when  $C_x$  is known as a result of some previous independent determination of  $x$ . BOSSLER [1972] has made a study of these two approaches and it is the main steps of his study that we are following here.

Common to both the Bayesian and generalized estimation methods is the a priori presence of the two covariance matrices  $C_x$  and  $C_l$ . Whether or not these are properly scaled, by  $\sigma_{0,x}^2$  and  $\sigma_{0,l}^2$ , respectively, is the crucial issue. Four different cases may arise (see TABLE 2):

- (a)  $\sigma_{0,x}^2$  and  $\sigma_{0,l}^2$  are both known. They may, naturally, be equal or different.
- (b)  $\sigma_{0,x}^2$  and  $\sigma_{0,l}^2$  are both unknown but equal; thus only one common factor is to be solved for.
- (c)  $\sigma_{0,x}^2$  is known and  $\sigma_{0,l}^2$  unknown, or vice versa. Again, only one unknown factor is solved for.
- (d)  $\sigma_{0,x}^2$  and  $\sigma_{0,l}^2$  are both unknown and suspected of being different. This case calls for a solution for two unknowns.

Let us first examine the solution to each of these four cases through the generalized adjustment approach, and follow with a similar examination from the Bayesian point of view.

In case (a) of the *generalized adjustment*, there are at least two techniques for obtaining the solution, and it is instructive to have a look at both. The first technique considers the mathematical model to be of the condition form (see (10.6)),

$$\boxed{f(l') = \mathbf{0}}, \quad (14.83)$$

where the hypervector  $l'$  is understood to be

$$l' = [l^T \mid x^{(0)T}]^T. \quad (14.84)$$

The linearized model is given by (12.2) in which the term  $A\delta$  is considered to be equal to zero, and where  $r$  is replaced by  $r' = [r^T \mid \delta^T]^T$ , and  $B$  by  $B'$  such that

$$B' = [B \mid A]. \quad (14.85)$$

Assuming that the cross-covariance matrix  $C_{lx} = \mathbf{0}$ , i.e.,  $l$  and  $x^{(0)}$  (the 0th approxi-

TABLE 14.2  
Summary of solutions by generalized adjustment approach

A priori knowledge			Formulae used		
Case	$\sigma_{0,i}^2$	Possible solutions	Corrections to parameters	Covariance matrix $C_x$	Estimated Variance Factor
Standard	Unknown assumption $C_x^{-1} = \mathbf{0}$	Improper adjustment	$\hat{\delta} = -\tilde{N}^{-1} \mathbf{A}^T \tilde{M} \mathbf{w}$	$\hat{C}_x = \hat{\sigma}_{0,x}^2 \tilde{N}^{-1}$	$\hat{\sigma}_{0,x}^2 = \frac{\hat{P}^T \hat{P}}{m-u}$
a	Known (common)	Adjustment condition model or implicit model	$\hat{\delta} = -(N + C_x^{-1})^{-1}$ $\times \mathbf{A}^T \mathbf{M} \mathbf{w}$	$\hat{C}_x = (N + C_x^{-1})^{-1}$	Known
b	Unknown (common)	Adjustment	Same as above except for tildes <sup>a</sup> , and $C_x = P_x^{-1}$		
c	Unknown Known	See text	$\hat{\delta} = -(\hat{\sigma}_{0,i}^{-2} \tilde{N} + C_x^{-1})^{-1}$ $\times \hat{\sigma}_{0,i}^{-2} \mathbf{A}^T \tilde{M} \mathbf{w}$	$\hat{C}_x = (\hat{\sigma}_{0,i}^{-2} \tilde{N} + C_x^{-1})^{-1}$	$\hat{\sigma}_{0,i}^2 = \frac{\hat{P}^T \hat{P}}{n}$
d	Unknown Unknown (different)	Adjustment	$\hat{\delta} = -(\hat{\sigma}_{0,i}^{-2} \tilde{N} + \hat{\sigma}_{0,x}^{-2} P_x)^{-1}$ $\times \hat{\sigma}_{0,i}^{-2} \mathbf{A}^T \tilde{M} \mathbf{w}$	$\hat{C}_x = (\hat{\sigma}_{0,i}^{-2} \tilde{N} + \hat{\sigma}_{0,x}^{-2} P_x)^{-1}$	$\hat{\sigma}_{0,i}^2 = \frac{\hat{P}^T \hat{P}}{n-u}; \hat{\sigma}_{0,x}^2 = \frac{\hat{\delta}^T \hat{P}_x \hat{\delta}}{n}$

<sup>a</sup> Note: matrices ( $M$  and  $N$ ) denoted by a tilde are misscaled, i.e.,  $P_i^{-1}$  is used instead of  $C_i$ .

mation of the parameters) are statistically independent, the corresponding covariance matrix of  $\mathbf{l}'$  is

$$\mathbf{C}'_l = \begin{bmatrix} \mathbf{C}_l & \mathbf{0} \\ \mathbf{0} & -\mathbf{C}_x \end{bmatrix}. \quad (14.86)$$

Substituting the primed quantities for the corresponding quantities in (12.26) and (12.36), one obtains the formulae given in TABLE 2. A comparison of these formulae with the solution of the standard adjustment problem, recapitulated for convenience in the same table, convinces us that the standard adjustment has the assumption of  $\mathbf{C}_x^{-1} = \mathbf{0}$  built in, as claimed at the beginning of this section.

The second technique, of including a known a priori  $\mathbf{C}_x^{-1}$ , uses the following two models:

$$\mathbf{f}_l(\mathbf{x}, \mathbf{l}) = \mathbf{0}, \mathbf{C}_l, \quad (14.87)$$

$$\mathbf{f}_x(\mathbf{x}, \mathbf{x}^{(0)}) = \mathbf{x} - \mathbf{x}^{(0)} = \mathbf{0}, \mathbf{C}_x. \quad (14.88)$$

The first model represents the situation characterized by the absence of the a priori knowledge of  $\mathbf{C}_x$ , while the second represents the situation when the a priori knowledge  $\mathbf{x}^{(0)}$  of  $\mathbf{x}$  is expressed in terms of the corresponding covariance matrix  $\mathbf{C}_x$ . As the reader can derive for himself, the linearized mathematical models are

$$\begin{bmatrix} \mathbf{B} & \mathbf{A} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \boldsymbol{\delta} \end{bmatrix} + \begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad (14.89)$$

and the corresponding covariance matrix has already been given by (86). The solution is again obtained directly from the equations in Chapter 12 simply by using the above redefined matrices and vectors. The resulting formulae are indeed identical with those obtained from the first technique (row (a) of TABLE 2).

Case (b) differs from case (a) as much as the case of the unknown variance factor  $\sigma_0^2$  differs from the case of the known variance factor in the standard adjustment approach. The difference is that  $\sigma_0^2$  has to be estimated a posteriori by  $\hat{\sigma}_0^2$ , and instead of getting  $\mathbf{C}_{\hat{x}}$  one gets  $\hat{\mathbf{C}}_{\hat{x}}$ . The estimate is derived from (12.51),

$$\hat{\sigma}_0^2 = \frac{\hat{\mathbf{r}}^T \mathbf{P}_l \hat{\mathbf{r}} + \hat{\boldsymbol{\delta}}^T \mathbf{P}_x \hat{\boldsymbol{\delta}}}{v}, \quad (14.90)$$

where the redundancy is

$$v = m - u = (n + u) - u = n. \quad (14.91)$$

The solution of case (c) parallels that of case (b) with the only exception being that  $\mathbf{C}_x^{-1}$  does not have to be scaled. The scale of  $\mathbf{P}_l$ , i.e.,  $\sigma_{0,l}^2$ , is estimated by means of (12.51), using only the model  $\mathbf{f}_l$  and  $v = n - u$ . If  $\mathbf{C}_l^{-1}$  is known and  $\sigma_{0,x}^2$  is to be estimated, then the model  $\mathbf{f}_x$  with the same  $v = n - u$  is used. The expression for the covariance matrix  $\hat{\mathbf{C}}_{\hat{x}}$ , as given in TABLE 2, is only approximate; THEIL [1963] has shown that the uncertainty of this formula is of the order of  $1/n$ . THEIL [1963] has

also proposed a second solution applicable when the a priori variances and covariances for the parameters are approaching their limits of 0 and  $\infty$ , which, in some cases, results in ill-conditioned models.

An exact solution of case (d) is not possible. An estimation, based on the notion of unbiasedness was devised by GRAFARENDS AND D'HONE [1978]. The solution shown in TABLE 2 is only approximate.

Before dealing with Bayesian ideas, let us have a look at the situation when there is a priori information available for only some of the parameters. Let us designate this part of  $x$  by  $x_2$  and its covariance matrix by  $C_{x_2}$ , understanding  $x_1$  to be those parameters for which there is no a priori information, i.e.,  $C_{x_1}^{-1} = \mathbf{0}$ . The normal equations arising from this situation follow from the normal equations (12.25) by analogy with row (a) in TABLE 2; namely,

$$\begin{bmatrix} N_1 & | & N_{12} \\ \hline N_{21} & | & N_{22} + C_{x_2}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\delta}_1 \\ \hline \hat{\delta}_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ \hline u_2 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \hline \mathbf{0} \end{bmatrix}. \quad (14.92)$$

Note that the a priori information, in the form of  $C_{x_2}$ , simply enters into the mathematical model as a contribution coming from observations. More about this point will be said in §14.5 and §14.6.

It has long been recognized that an a priori belief in some property of a parameter may play a significant role in the parameter's determination [WONNACOTT AND WOONACOTT, 1972]. Essentially, the *Bayesian approach* formalizes this idea by quantifying these subjective beliefs. The explicit use of probability density functions for the various quantities is a distinguishing feature of this approach. The three probability density functions involved are:  $\phi_x(\xi; x^{(0)}, C_x)$ ,  $\phi_{I/x}(\xi; I, C_I)$ , and  $\phi_{x/I}(\xi; \hat{x}, C_{\hat{x}})$ . The first probability density function is merely the expression of the a priori statistical information about the parameters  $x$ . The second probability density function, also called the *likelihood function*, describes the conditional probability of  $I$ , given the knowledge that  $x$  has occurred (cf. §3.4). Loosely interpreted, this simply means that the observations have a bearing on the determination of the parameters. Again, the form of this probability density function must be specified; usually the multivariate normal (see (13.25)) is postulated along with its parameters  $I$  and  $C_I$ . The third probability density function is the unknown a posteriori probability density function, i.e., the stochastical representation of the parameters we are seeking.

Bayes's theorem [WONNACOTT AND WOONACOTT, 1972]:

$$\phi_{x/I}(\xi; \hat{x}, C_{\hat{x}}) = \frac{\phi_{I/x}(\xi; I, C_I) \phi_x(\xi; x^{(0)}, C_x)}{\psi}, \quad (14.93)$$

is the mathematical instrument used to solve for the unknown probability density function  $\phi_{x/I}(\xi; \hat{x}, C_{\hat{x}})$  along with its parameters  $\hat{x}$  and  $C_{\hat{x}}$ . The real number  $\psi$  acts simply as a normalizing factor and has little significance from the conceptual point of view. Note that the Bayesian approach not only seeks to determine the  $\hat{x}$  and  $C_{\hat{x}}$  but, unlike the generalized adjustment approach, also determines the probability density function of the parameters.

BOSSLER [1972], using Bayes's theorem, has demonstrated that the case of no a priori information about the parameters ( $C_x^{-1} = \mathbf{0}$ ) and known variance factor  $\sigma_{0,1}^2$ , gives results identical to those of the standard adjustments (cf. TABLE 2). He has also shown that, for the other four cases, the Bayesian results for  $\hat{x}$  and  $C_{\hat{x}}$  are identical with the generalized adjustments approach. For the cases of unknown variance factor  $\sigma_0^2$ , however, the Bayesian estimate  $\sigma_0^2$  is  $(m - u)/(m - u - 2)$  times that derived in (12.51). Thus, the Bayesian estimate differs from the standard estimate by a factor of 2 in the denominator. The Bayesian estimate is thus consistent with the general principle that one degree of freedom is lost for each unknown parameter of the probability density function that is being estimated. In this instance, the 2 corresponds to the two parameters of the postulated, multivariate, normal probability density function of the parameters—the mean and covariance matrix. Clearly, for a large degree of freedom, this difference is not significant.

It should be noted that even when using the Bayesian approach, it is still possible to fully exploit the statistical developments of Chapter 13. The only change that occurs concerns the degrees of freedom which become smaller by two. The consequence of this fact is that confidence intervals for the same confidence level become larger. This result is intuitively pleasing; the Bayesian approach is, after all, based on subjective beliefs, and one should thus expect the results to be less trustworthy compared with results based on some objective criteria.

#### 14.5. Problems with constraints and singularities

In practice, one often encounters problems somewhat more complex than those treated so far. These problems are of two different kinds: those where more is known about the unknown parameters than expressed in the main mathematical model  $f$ ; and those that, after the formulation of the mathematical model, still yield no solution because of singularities. Common to both kinds of problems is the concept of *constraints*. The existence of additional information for the first category of problems can be understood as the existence of constraints on the main mathematical model, while at least some singular problems can also be solved by imposing constraints. Let us begin with problems of the first kind.

The formulation of constraints arising from additional information about the unknown parameters amounts to a formulation of an additional mathematical model which will be assumed here to contain only the unknown parameters. Thus, a problem with constraints can be formulated in terms of two models:

$$\boxed{f(x, t) = 0,} \quad (14.94)$$

$$\boxed{f_c(x) = 0.} \quad (14.95)$$

It will be assumed that it is possible to solve for  $x$  using only the main model  $f$ . The auxiliary model  $f_c$  consists of  $m_c$  constraint functions that reflect some mathematical or physical laws. Such constraints are known as *absolute constraints*. (Another

variety, known as *weighted constraints*, are those based on observed rather than theoretical relations among the parameters and, as such, usually contain the observations as well. The latter constraints can be trusted only in a limited sense and thus have some finite weights associated with them. These will not be treated here, and the interested reader is referred to SCHWARZ [1969] and MIKHAIL [1976].) The above models are next linearized to yield (see FIG. 10)

$$\mathbf{A}\delta + \mathbf{B}r + w = \mathbf{0}, \quad (14.96)$$

$$\mathbf{D}\delta + w_c = \mathbf{0}. \quad (14.97)$$

Then the variation function for finding the least-squares solution is written, similar to (12.12), as

$$\phi = r^T C_r^{-1} r + 2k^T (\mathbf{A}\delta + \mathbf{B}r + w) + 2k_c^T (\mathbf{D}\delta + w_c), \quad (14.98)$$

where  $\mathbf{C}_r \equiv \mathbf{C}_l$  is the covariance matrix of the observations. This time, however, we have two sets of Lagrange correlates,  $k$ ,  $k_c$ , reflecting the fact that two models are present. The minimum with respect to  $r$  is found the same way as in §12.2. After eliminating the residual vector  $\hat{r}$  and the first set of correlates  $k$ , the equation system

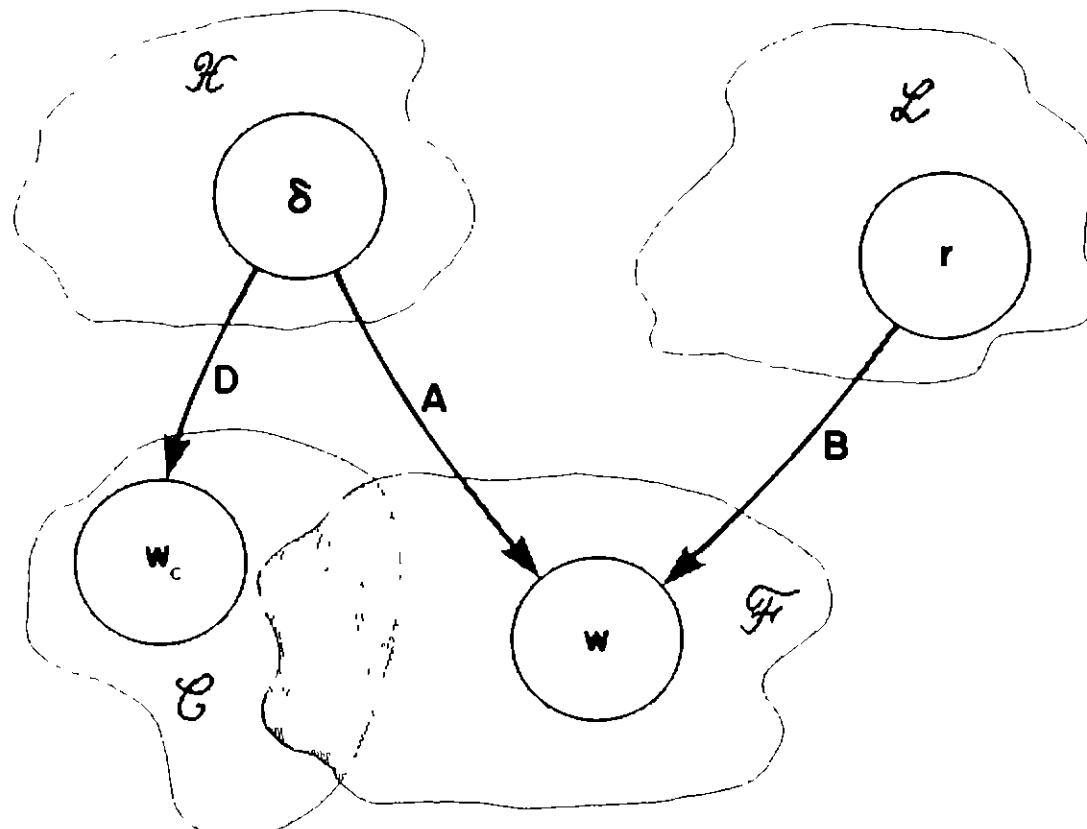


FIG. 14.10. Spaces used for solving constrained problems.

describing the combination of  $f$  and  $f_c$  becomes

$$\begin{bmatrix} \mathbf{N}^{-1} & \mathbf{D}^T \\ \mathbf{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta} \\ \hat{\mathbf{k}}_c \end{bmatrix} + \begin{bmatrix} \mathbf{u} \\ \mathbf{w}_c \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad (14.99)$$

where  $\mathbf{N}$  is given by (12.23) and  $\mathbf{u}$  by (12.24).

Since a solution from  $f$  alone is stipulated possible, it means that  $\mathbf{N}$  is required to be regular and  $\boldsymbol{\delta}$  may also be eliminated from (99). The remaining equation yields

$$\hat{\mathbf{k}}_c = (\mathbf{D}\mathbf{N}^{-1}\mathbf{D}^T)^{-1}(\mathbf{w}_c - \mathbf{D}\mathbf{N}^{-1}\mathbf{u}). \quad (14.100)$$

From (99), one also gets

$$\mathbf{N}\boldsymbol{\delta} + \mathbf{D}^T\hat{\mathbf{k}}_c + \mathbf{u} = \mathbf{0}, \quad (14.101)$$

and, substituting for  $\hat{\mathbf{k}}_c$  from (100), the above equation gives

$$\hat{\boldsymbol{\delta}} = \boldsymbol{\delta}^{(1)} - \mathbf{N}^{-1}\mathbf{D}^T(\mathbf{D}\mathbf{N}^{-1}\mathbf{D}^T)^{-1}(\mathbf{w}_c + \mathbf{D}\boldsymbol{\delta}^{(1)}), \quad (14.102)$$

where

$$\boldsymbol{\delta}^{(1)} = -\mathbf{N}^{-1}\mathbf{u} \quad (14.103)$$

represents the solution from the main model  $f$  alone. Note that the corrective term  $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^{(1)}$  arises from the enforcement of the constraints.

Back substitution yields the solutions for the other vectors. Of particular interest are the solutions for the residual vector  $\mathbf{r}^{(1)}$ , describing the fit of  $\mathbf{l}$  to  $f$  alone, and  $\hat{\mathbf{r}}$ , corresponding to the combination of  $f$  and  $f_c$ . Both results are obtained from (12.29) when  $\boldsymbol{\delta}^{(1)}$  and  $\hat{\boldsymbol{\delta}}$  are used respectively. The covariance matrix of  $\mathbf{r}^{(1)}$  is given by (12.38), while the derivation of that for  $\hat{\mathbf{r}}$  is left as an exercise for the reader.

The next task is to obtain the covariance matrix of the parameters. Equation (102) may be written in a form,

$$\hat{\boldsymbol{\delta}} = -(\mathbf{N}^{-1} - \mathbf{N}^{-1}\mathbf{D}^T(\mathbf{D}\mathbf{N}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{N}^{-1})\mathbf{A}^T\mathbf{M}\mathbf{w} + \text{const.}, \quad (14.104)$$

where the first term is a function of observations (through  $\mathbf{w}$ ), while the second term is constant with respect to  $\mathbf{l}$ . By recognizing that

$$\mathbf{C}_w = \mathbf{M}^{-1}, \quad \mathbf{D}\mathbf{N}^{-1}\mathbf{A}^T\mathbf{M}\mathbf{A} = \mathbf{D}, \quad \mathbf{D}\mathbf{N}^{-1}\mathbf{A}^T\mathbf{M}(\mathbf{D}\mathbf{N}^{-1}\mathbf{A}^T)^T = \mathbf{D}\mathbf{N}^{-1}\mathbf{D}^T,$$

one can see that the covariance law applied to (104) yields

$$\mathbf{C}_{\hat{\boldsymbol{\delta}}} = \mathbf{N}^{-1} - \mathbf{N}^{-1}\mathbf{D}^T(\mathbf{D}\mathbf{N}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{N}^{-1}.$$

Denoting  $\mathbf{N}^{-1}$  by  $\mathbf{C}_{\boldsymbol{\delta}}^{(1)}$ , one finally gets

$$\mathbf{C}_{\hat{\boldsymbol{\delta}}} = \mathbf{C}_{\boldsymbol{\delta}}^{(1)} - \mathbf{C}_{\boldsymbol{\delta}}^{(1)}\mathbf{D}^T(\mathbf{D}\mathbf{C}_{\boldsymbol{\delta}}^{(1)}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{C}_{\boldsymbol{\delta}}^{(1)}. \quad (14.105)$$

In the situation when the scale of  $C_i$ , i.e.,  $\sigma_0^2$ , is unknown, its estimate is obtained from (12.51) with degrees of freedom given as

$$v = (m + m_c) - u. \quad (14.106)$$

An interpretation of the main results ((102) and (105)) is now in order. It is clear that what we have are sequential expressions: to compute the solution vector  $\hat{\delta}$  (or its covariance matrix  $C_{\hat{\delta}}$ ), one simply computes the solution  $\delta^{(1)}$  (or  $C_{\delta}^{(1)}$ ) corresponding to  $f$  alone and then subtracts the corrective term. This important concept will be treated systematically in §14.6.

Let us now turn to the *problems with singularities*. As stated above, one can positively identify a singularity only after the mathematical model is available, so, strictly speaking, we should speak about *singular mathematical models*. Singularity of models is caused by one or both of the following reasons: the problem is ill-posed, whereby the wrong questions are posed or the wrong answers sought; the model is improperly formulated, whereby dependent equations or too many independent parameters are included. The problem may be ill-posed because we expect too much from the available observations, too few observables were observed, or they are of the wrong kind. Be that as it may, it is sometimes difficult to anticipate that the problem will turn out to be ill-posed. The faulty formulation of the model may turn out to be equally difficult to detect.

For these reasons, one wants to have a diagnostic mechanism capable of pointing out the singularity as early as possible in the attempted solution. The usual *diagnosis of singularity* is done by investigating the rank deficiency of either the design matrix  $A$  of the linearized model or of the matrix of normal equations  $N$ . If

$$\text{rank } A = \text{rank } N < u, \quad (14.107)$$

one gets  $\det N = 0$ , and  $N$  is singular. There exists an infinite set of solutions of such a (linearized) mathematical model, and the mathematical model is singular. One class of singular models was shown already in §11.3, along with one possible treatment. It should be pointed out that even non-singular models may have a very small value for the determinant of  $N$ . This points out the ill-conditioning of  $N$  (see §3.1) which may virtually prevent one from getting a meaningful solution. Singularity is indeed merely the limiting case of ill-conditioning, and thus the techniques of dealing with singular problems can be applied to ill-conditioned problems as well.

Once a singularity is discovered, a return to the problem formulation stage is necessary to ensure that the model is assembled correctly. If an erroneous formulation of the model can be ruled out as the cause of the singularity (or ill-conditioning), it can be assumed that the problem is ill-posed. All the reasons for a problem to be ill-posed can be reduced to only one: namely, the wrong kind of parameters are sought. From the knowledge of the problem, one should be able to identify the *indeterminable parameters* or, at least, the cause of their *indeterminacy*. For instance, if only angles in a triangle are observed, the sides are indeterminable; more involved situations are cited in Part IV, and §27.2. Let us now denote the defect (see §3.1) of  $N$  by  $d$ ; i.e.,

$$d = \text{def } N = u - \text{rank } N. \quad (14.108)$$

Then, after the cause of singularity is known, an infinite set of solutions to the problem should be obtainable in one of the following two ways [THOMPSON, 1969]: by expressing the first  $u - d$  unknowns as functions of the remaining  $d$  unknowns; or by expressing all the unknowns as functions of  $d$  common quantities known to be responsible for the indeterminacy. Taking the above example further, two sides of the triangle may be expressed either as functions of the unspecified, variable third, or all three sides may be expressed as functions of a common, unspecified, variable scale factor.

Another way of solving a singular problem is by imposing a particular set of constraints. These constraints, whatever they may be, can be written in the same way as the constraints at the beginning of this section: i.e., as (95) or, in linearized form, as (97). Hence the rank deficient design matrix  $\mathbf{A}$  can be augmented in the following manner

$$\mathbf{A} \rightarrow \begin{bmatrix} \mathbf{A} \\ \mathbf{D} \end{bmatrix}, \quad (14.109)$$

or, similarly, the matrix of normal equations can become

$$\mathbf{N} \rightarrow \begin{bmatrix} \mathbf{N} & \mathbf{D}^T \\ \mathbf{D} & \mathbf{0} \end{bmatrix}. \quad (14.110)$$

It can be stated that for a solution to exist, the rank of the augmented matrices must fulfil the following inequality:

$$\text{rank} \begin{bmatrix} \mathbf{A} \\ \mathbf{D} \end{bmatrix} \equiv \text{rank} \begin{bmatrix} \mathbf{N} & \mathbf{D}^T \\ \mathbf{D} & \mathbf{0} \end{bmatrix} \geq u. \quad (14.111)$$

If the rank is greater than  $u$ , then the problem is said to have an *overconstrained solution*; this is the situation treated at the beginning of this section. It should be realized that in an overdetermined solution it is quite possible for rank  $\mathbf{D}$  to be less than the number of constraint equations and still be greater than the rank defect in  $\mathbf{A}$ ; this situation could arise if some dependent constraints are included.

Let us now look at the situation in which rank  $\mathbf{D}$  equals the number of constraint equations and this, in turn, equals the rank defect of  $\mathbf{A}$ . Then, one has

$$\text{rank} \begin{bmatrix} \mathbf{A} \\ \mathbf{D}_m \end{bmatrix} = \text{rank } \mathbf{A} + \text{rank } \mathbf{D}_m = (u - d) + d = u.$$

(14.112)

Constraints satisfying the above equation are called *minimal constraints*, and their design matrix is denoted by  $\mathbf{D}_m$ . These give what is known as a *minimum constraint solution*. This solution is obtained by following the same general principles as those used in the first part of this section. The only difference, as will be seen later, is that a special algorithm is needed to avoid the singularity inherent in  $\mathbf{N}$ .

A particularly popular kind of minimal constraints is the *inner constraints*, which are those minimal constraints fulfilling the following condition:

$$\mathbf{A} \mathbf{D}_i^T = \mathbf{0}. \quad (14.113)$$

Abstract as it may seem, it is possible to prove [BLAHA, 1982] that this condition leads to the following property:

$$\min_{\mathbf{D}_i \in \mathbf{D}_m} \text{tr } \mathbf{C}_{\delta}, \quad (14.114)$$

which means that this particular choice of  $\mathbf{D}_m$  generally leads to smaller variances of the parameters as a whole. Formally, the property given by (114) constitutes the definition of the inner constraint design matrix  $\mathbf{D}_i$ . The fact that the inner constraints conform to the general behaviour of least-squares solutions (cf. property (c) in §13.1) is, of course, their main attraction.

It is useful to point out that an entire family of constraints  $\mathbf{H}_i$  can be generated from  $\mathbf{D}_i$  simply through multiplication by some matrix  $\mathbf{J}$ : namely,

$$\mathbf{H}_i = \mathbf{J} \mathbf{D}_i, \quad (14.115)$$

where  $\mathbf{J}$  must be conformable with  $\mathbf{D}_i$  and regular. In this context,  $\mathbf{D}_i$  is said to represent the set of basic inner constraints for which a specific example will be given in §17.1. In practice,  $\mathbf{J}$  is usually chosen as an identity matrix. In some applications, the multiplication matrix  $\mathbf{J}$  has been chosen as

$$\mathbf{J} = \mathbf{D}_a^{-1}, \quad (14.116)$$

where  $\mathbf{D}_a$  is the first  $d$  by  $d$  submatrix of the partitioned  $d$  by  $u$  matrix of inner constraints

$$\mathbf{D}_i = [\mathbf{D}_a \mid \mathbf{D}_b]. \quad (14.117)$$

BLAHA [1982] utilized this form in showing the equivalence between the inner constraint method of PERELMUTER [1979] and the standard inner constraint method in which  $\mathbf{J} = \mathbf{I}$ .

The constrained singular case using inner constraints can be written as (cf. (99))

$$\begin{bmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\mathbf{k}}_c \end{bmatrix} = - \begin{bmatrix} \mathbf{N}^{-1} \mathbf{D}_i^T \\ \mathbf{D}_i \mid \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{u} \\ \mathbf{w}_c \end{bmatrix}, \quad (14.118)$$

where  $\mathbf{w}_c$  is usually set to a null vector. This selection does not affect the property of  $\min \text{tr } \mathbf{C}_{\delta}$  but simply biases the solution in a certain way; for details, the interested reader is referred to BOSSLER ET AL. [1973]. Returning to the problem at hand, it can be seen that (118) cannot be solved in the same way as (99) because  $\mathbf{N}$  is now singular. The solution must be sought using the following substitution (see, e.g.,

THOMPSON [1969]):

$$\begin{bmatrix} \mathbf{N}^+ & \mathbf{D}_i^T \\ \mathbf{D}_i & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{R} & \mathbf{D}_i^T(\mathbf{D}_i \mathbf{D}_i^T)^{-1} \\ (\mathbf{D}_i \mathbf{D}_i^T)^{-1} \mathbf{D}_i & \mathbf{0} \end{bmatrix}, \quad (14.119)$$

where

$$\mathbf{R} = [\mathbf{N} + \mathbf{K} \mathbf{D}_i^T(\mathbf{D}_i \mathbf{D}_i^T)^{-1} \mathbf{D}_i]^{-1} [\mathbf{I} - \mathbf{D}_i^T(\mathbf{D}_i \mathbf{D}_i^T)^{-1} \mathbf{D}_i], \quad (14.120)$$

and  $\mathbf{K}$  is an arbitrary, non-singular matrix. From (118) and (119),

$$\hat{\boldsymbol{\delta}} = -\mathbf{R}\mathbf{u}. \quad (14.121)$$

BLAHA [1971] has proved that, in addition,

$$\mathbf{C}_{\hat{\boldsymbol{\delta}}} = \mathbf{R}, \quad (14.122)$$

where  $\mathbf{R}$  is symmetrical and unique in spite of the introduction of the arbitrary matrix  $\mathbf{K}$ . It is normally expedient to select  $\mathbf{K}$  such that  $\mathbf{K} = \mathbf{K}\mathbf{I}$  ( $K > 0$ ) or, even more simply,  $\mathbf{K} = \mathbf{I}$  (i.e.,  $K = 1$ ). For special constraints, other appropriate selections of  $\mathbf{K}$  can be made.

The solution of the singular constrained problem (118) can be approached in yet another way—that of using *generalized matrix inverses*. The reason for choosing this route is, again, the singularity of  $\mathbf{N}$ . The generalized inverse  $\mathbf{S}^-$ , sometimes for brevity called a *g-inverse*, of a singular rectangular matrix  $\mathbf{S}$  is defined as

$$\mathbf{S}^- \Leftrightarrow \mathbf{S}\mathbf{S}^-\mathbf{S} = \mathbf{S}. \quad (14.123)$$

RAO AND MITRA [1971] list a variety of possible g-inverses, from which at least the *minimum norm g-inverse* ( $\mathbf{S}_m^-$ ) and *least-squares g-inverse* ( $\mathbf{S}_t^-$ ) should be mentioned here. The most restricted of the g-inverses is the already mentioned (§11.3) pseudo-inverse, also known as the *Moore–Penrose g-inverse* ( $\mathbf{S}^+$ ), defined as

$$\mathbf{S}^+ \Leftrightarrow \mathbf{S}\mathbf{S}^+\mathbf{S} = \mathbf{S} \quad \text{and} \quad \mathbf{S}^+\mathbf{S}\mathbf{S}^+ = \mathbf{S}^+. \quad (14.124)$$

In fact, it can be shown that the above introduced matrix  $\mathbf{R}$  is a pseudo-inverse of  $\mathbf{N}$ : i.e.,

$$\mathbf{R} = \mathbf{N}^+. \quad (14.125)$$

Accordingly, the pseudo-inverse solution yields equations identical with those given above ((121) and (122)). As already quoted in §11.3, the pseudo-inverse solution  $\hat{\boldsymbol{\delta}}_p$  has the property of  $\min_{\hat{\boldsymbol{\delta}}} \hat{\boldsymbol{\delta}}^T \hat{\boldsymbol{\delta}} = \min_{\hat{\boldsymbol{\delta}}} \|\hat{\boldsymbol{\delta}}\|$  because the pseudo-inverse is also a minimum norm g-inverse. This is in addition to the above discussed property of  $\min_{\hat{\boldsymbol{\delta}}} \text{tr } \mathbf{C}_{\hat{\boldsymbol{\delta}}}$  (114) and the obvious property of  $\min_{\hat{\boldsymbol{\delta}}} \|\hat{\mathbf{r}}\|$ ; the pseudo-inverse is also a least-squares g-inverse. Note that both these norms are in different spaces and thus

taken with respect to different metrics. The metric in  $\mathcal{X}$  is  $I$ , in  $\mathcal{L}$  it is  $C_r^{-1}$ . Other mathematically equivalent methods of implementing inner constraints to solve singular problems exist.

#### 14.6. Step-by-step procedures in dynamic and static problems

The general transitional, or step-by-step, procedure was described in §10.3 by (10.23) where  $S$  denoted the transition matrix from one state to another state and  $s$  denoted the state vector. Generally,  $s$  may be static (see FIG. 11) or may change from one step to another. It is useful to think of the first case as being a general iterative process. In the time-varying case, the state vector is given subscripts  $k-1, k, k+1, \dots$

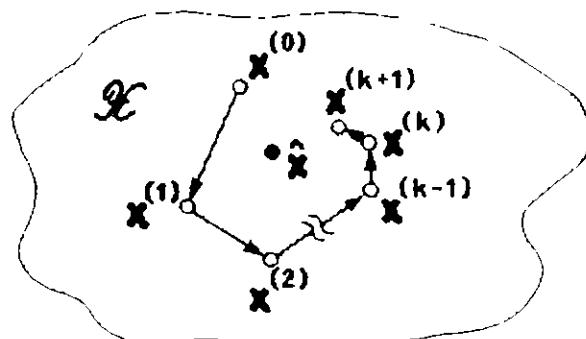


FIG. 14.11. Sequential (iterative) process with static state vector  $x$ .

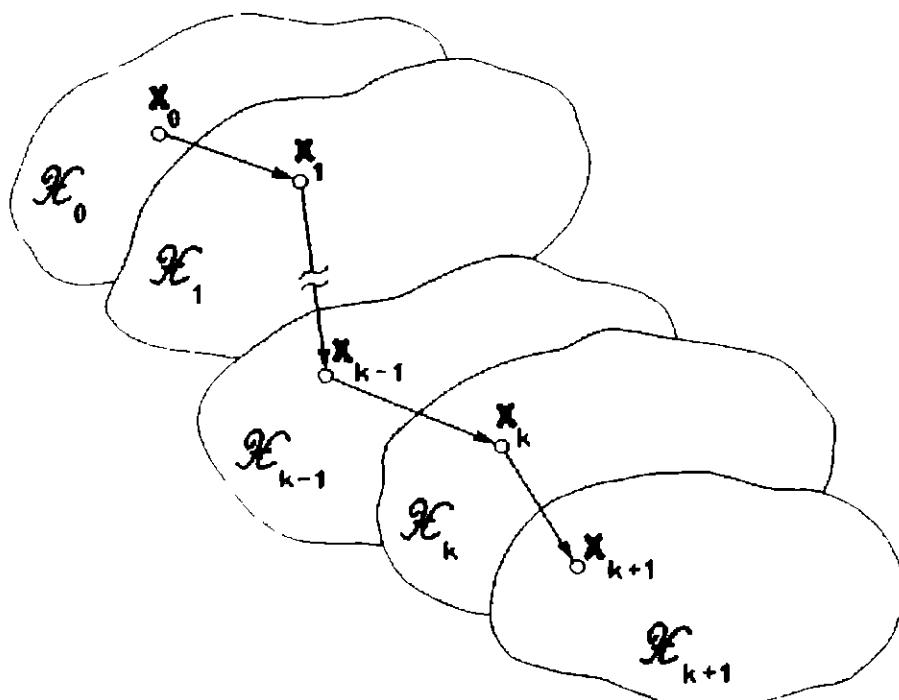


FIG. 14.12. Sequential process with changing state vector  $x$ .

denoting steps in time. These subscripts then generally correspond to epochs  $\tau_{k-1}, \tau_k, \tau_{k+1}, \dots$ , and each state vector belongs to a different solution space (see FIG. 12).

An example of the step-by-step procedure for the static situation, where the subscripts do not refer to time epochs, was encountered in §14.5 where (102) could have been written as

$$\hat{\delta}^{(k)} = S_{k,k-1} \delta^{(k-1)} + \text{const.} \quad \text{for } k=2. \quad (14.126)$$

There, the transition matrix  $S_{k,k-1}$  had a specific form in terms of the inverse  $N^{-1}$  and the constraint design matrix  $D$ , and the constant term was also a function of  $D$ ,  $N$ , and  $w_c$ . The equation shows the application of only one constraint but it is not difficult to see that successive constraints could be applied, one after the other, making it an iterative process. Similarly, the iterative process for removing the effect of non-linearity discussed in §12.2 can be thought of as a step-by-step procedure of the static variety. Another example of the step-by-step procedure was encountered in §14.4, where the combination of two mathematical models ((87) and (88)) could have been written as

$$f_{k-1}(x_{k-1}, l_{k-1}) = \mathbf{0}, \quad f_k(x_k, l_k) = \mathbf{0}. \quad (14.127)$$

There, two different sets of ‘observations’,  $l_{k-1} = l$  and  $l_k = x^{(0)}$ , were used to estimate the same unknown parameters  $\hat{x}$ . The solution  $\hat{x}$  was not written in a sequential form either; this will be done in this section as an application of a more general time-varying, or dynamic, case which is to be discussed next.

It is gratifying to know that all the step-by-step procedures discussed in the geodetic literature are simply special cases of the most complex case known as *Kalman filtering*. Below an overview of the development of the *Kalman filter* equations is given using the concepts presented in Chapters 12 and 14. The reader can find alternative derivations in MORRISON [1969] and MORITZ [1973]; the original derivation given in KALMAN [1960] is more complicated. To begin with, the parameters  $x$  as well as observations  $l$  are regarded as varying with time. Denoting the present epoch by  $\tau_k$  and the previous epoch by  $\tau_{k-1}$ , the three models constituting the filter are written as

$$f(x_{k-1}, l_{k-1}) = \mathbf{0} \quad \text{for } \tau_{k-1}, \quad (14.128)$$

$$f(x_k, l_k) = \mathbf{0} \quad \text{for } \tau_k, \quad (14.129)$$

$$f_{k-1,k}(x_{k-1}, x_k) = \mathbf{0} \quad \text{for } \langle \tau_{k-1}, \tau_k \rangle, \quad (14.130)$$

where the first two are the primary models, taken here as being of the same form  $f$ , at the two epochs  $\tau_k$  and  $\tau_{k-1}$ . The function  $f_{k-1,k}$  is the secondary, or *dynamic, model* giving the functional relationship (constraint) between the state vectors at successive epochs  $\tau_{k-1}, \tau_k$ . This dynamic model is simply our old acquaintance from §10.3—the model of a dynamic process. The two observation vectors  $l_{k-1}, l_k$  are considered mutually statistically independent, with regular covariance matrices  $C_{l_{k-1}}$  and  $C_{l_k}$ ; thus the cross-covariance matrix  $C_{l_{k-1}l_k} = \mathbf{0}$ . Further, it is stipulated that  $\dim f_{k-1,k} = \dim x = u$ .

Clearly, the primary models have the following linear versions:

$$\mathbf{f}(\mathbf{x}_{k-1}, \mathbf{l}_{k-1}) = \mathbf{w}_{k-1} + \mathbf{A}_{k-1}\boldsymbol{\delta}_{k-1} + \mathbf{B}_{k-1}\mathbf{r}_{k-1} = \mathbf{0}, \quad (14.131)$$

$$\mathbf{f}(\mathbf{x}_k, \mathbf{l}_k) = \mathbf{w}_k + \mathbf{A}_k\boldsymbol{\delta}_k + \mathbf{B}_k\mathbf{r}_k = \mathbf{0}, \quad (14.132)$$

where

$$\mathbf{w} = \mathbf{f}(\mathbf{x}^{(0)}, \mathbf{l}), \quad (14.133)$$

$$\boldsymbol{\delta} = \mathbf{x} - \mathbf{x}^{(0)}. \quad (14.134)$$

After linearization, the dynamic model becomes

$$\mathbf{f}_{k-1,k} = \mathbf{u}_{k-1,k} + \frac{\partial \mathbf{f}_{k-1,k}}{\partial \mathbf{x}_{k-1}} \boldsymbol{\delta}_{k-1} + \frac{\partial \mathbf{f}_{k-1,k}}{\partial \mathbf{x}_k} \boldsymbol{\delta}_k = \mathbf{0}. \quad (14.135)$$

Denoting the Jacobian matrix  $\partial \mathbf{f} / \partial \mathbf{x}_k$  by  $\mathbf{D}_k$  and assuming it to be regular, we get

$$\boldsymbol{\delta}_k = -\mathbf{D}_k^{-1} \mathbf{D}_{k-1} \boldsymbol{\delta}_{k-1} + \mathbf{D}_k^{-1} \mathbf{u}_{k-1,k}, \quad (14.136)$$

or

$$\boldsymbol{\delta}_k = \mathbf{S}_{k-1,k} \boldsymbol{\delta}_{k-1} + \boldsymbol{\epsilon}_{k-1,k}. \quad (14.137)$$

Note that when an explicit dynamic model  $\mathbf{x}_k = \mathbf{g}(\mathbf{x}_{k-1})$  is considered instead of the implicit (130), which is usually the case, there is no need for any matrix inversion and thus for the assumption of regularity of  $\mathbf{D}_k$ . It is interesting to compare this equation with (10.23). With the realization that  $\boldsymbol{\delta}$  plays the role of  $s$  in §10.3 and that  $\mathbf{S}_{k-1,k}$  is a transition matrix (obtained from only the dynamic model), the two equations differ only by the vector

$$\boldsymbol{\epsilon}_{k-1,k} = \mathbf{D}_k^{-1} \mathbf{f}_{k-1,k}(\mathbf{x}_{k-1}^{(0)}, \mathbf{x}_k^{(0)}), \quad (14.138)$$

or, in the case of the explicit model, by

$$\boldsymbol{\epsilon}_{k-1,k} = \mathbf{g}(\mathbf{x}_{k-1}^{(0)}) - \mathbf{x}_k^{(0)}, \quad (14.139)$$

called the *dynamic model error*. Evidently, this error is a projection of the misclosure  $\mathbf{u}_{k-1,k}$  of the dynamic process into the parameter space  $\mathcal{X}$ . If the selection of the expansion points  $\mathbf{x}_k^{(0)}$  is random, then  $\boldsymbol{\epsilon}_{k-1,k}$  is random as well. It is usually assumed that when the expansion points are obtained from

$$\mathbf{x}_k^{(0)} = \mathbf{g}(\hat{\mathbf{x}}_{k-1}),$$

(14.140)

making

$$\boldsymbol{\epsilon}_{k-1,k} = \mathbf{g}(\mathbf{x}_{k-1}^{(0)}) - \mathbf{g}(\hat{\mathbf{x}}_{k-1}), \quad (14.141)$$

the mathematical expectation  $E(\boldsymbol{\epsilon})$  of which is zero with a regular a priori covariance matrix  $C_{\boldsymbol{\epsilon}_{k-1,k}}$ .

The equations that give the solution to the combination of the above three models are obtained in the usual manner by first constructing the variation function. Because there are three random vectors  $\mathbf{r}_{k-1}, \mathbf{r}_k, \boldsymbol{\epsilon}_{k-1,k}$ , whose norms should be

minimized, and three models, there are also three quadratic forms plus the three linearized models and three vectors of correlates present in the variation function  $\phi$ ; namely,

$$\begin{aligned}\phi = & \mathbf{r}_{k-1}^T \mathbf{C}_{l_{k-1}}^{-1} \mathbf{r}_{k-1} + \mathbf{r}_k^T \mathbf{C}_{l_k}^{-1} \mathbf{r}_k + \boldsymbol{\epsilon}_{k-1,k}^T \mathbf{C}_{\epsilon_{k-1,k}}^{-1} \boldsymbol{\epsilon}_{k-1,k} \\ & + 2\mathbf{k}_{k-1}^T (\mathbf{A}_{k-1} \boldsymbol{\delta}_{k-1} + \mathbf{B}_{k-1} \mathbf{r}_{k-1} + \mathbf{w}_{k-1}) + 2\mathbf{k}_k^T (\mathbf{A}_k \boldsymbol{\delta}_k + \mathbf{B}_k \mathbf{r}_k + \mathbf{w}_k) \\ & + 2\mathbf{k}_{k-1,k}^T (-\mathbf{S}_{k-1,k} \boldsymbol{\delta}_{k-1} + \boldsymbol{\delta}_k - \boldsymbol{\epsilon}_{k-1,k}).\end{aligned}\quad (14.142)$$

After taking the partial derivatives of the variation function with respect to the variates and equating them to null matrices, the least-squares normal equations in their most expanded form are obtained:

$$\left[ \begin{array}{cccc|ccc|c} \mathbf{C}_{l_{k-1}}^{-1} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{k-1}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{l_k}^{-1} & \mathbf{0} & \mathbf{0} & \mathbf{B}_k^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_{\epsilon_{k-1,k}}^{-1} & \mathbf{0} & \mathbf{0} & -\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{B}_{k-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_k \\ \mathbf{0} & \mathbf{0} & -\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{S}_{k-1,k} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{k-1}^T & \mathbf{0} & -\mathbf{S}_{k-1,k}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_k^T & \mathbf{I} & \mathbf{0} & \mathbf{0} \end{array} \right] \times \begin{bmatrix} \hat{\mathbf{r}}_{k-1} \\ \hat{\mathbf{r}}_k \\ \hat{\boldsymbol{\epsilon}}_{k-1,k} \\ \hat{\mathbf{k}}_{k-1} \\ \hat{\mathbf{k}}_k \\ \hat{\mathbf{k}}_{k-1,k} \\ \hat{\boldsymbol{\delta}}_{k-1} \\ \hat{\boldsymbol{\delta}}_k \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \frac{\mathbf{w}_{k-1}}{\mathbf{w}_k} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \mathbf{0}. \quad (14.143)$$

The first step in obtaining the Kalman filter expressions is to eliminate the first four vectors using the technique shown in §3.1. Denoting, again, the weight matrix of observations transformed to the model space  $\mathcal{F}$  by  $\mathbf{M}$ , and rearranging the obtained matrix equation, we get the following result:

$$\left[ \begin{array}{cc|cc|cc|c} \mathbf{N}_{k-1} & -\mathbf{S}_{k-1,k}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{S}_{k-1,k} & -\mathbf{C}_{\epsilon_{k-1,k}} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{A}_k^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_k & -\mathbf{M}_k^{-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array} \right] \begin{bmatrix} \boldsymbol{\delta}_{k-1} \\ \hat{\mathbf{k}}_{k-1,k} \\ \hat{\boldsymbol{\delta}}_k \\ \hat{\mathbf{k}}_k \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{k-1} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{w}_k \end{bmatrix} = \mathbf{0}. \quad (14.144)$$

Here, as usual (cf. (12.23)),

$$\mathbf{N}_{k-1} = \mathbf{A}_{k-1}^T (\mathbf{B}_{k-1} \mathbf{C}_{l_{k-1}} \mathbf{B}_{k-1}^T)^{-1} \mathbf{A}_{k-1} = \mathbf{A}_{k-1}^T \mathbf{M}_{k-1} \mathbf{A}_{k-1}, \quad (14.145)$$

and the new right-hand side is (cf. (12.24))

$$\mathbf{u}_{k-1} = \mathbf{A}_{k-1}^T \mathbf{M}_{k-1} \mathbf{w}_{k-1}. \quad (14.146)$$

The first equation of the above system yields the following solution for the increment of the state vector in  $(k-1)$ st step

$$\hat{\boldsymbol{\delta}}_{k-1} = \boldsymbol{\delta}_{k-1}^{(1)} + \mathbf{N}_{k-1}^{-1} \mathbf{S}_{k-1, k}^T \hat{\mathbf{k}}_{k-1, k}, \quad (14.147)$$

where  $\boldsymbol{\delta}_{k-1}^{(1)} = -\mathbf{N}_{k-1}^{-1} \mathbf{u}_{k-1}$  is the partial solution for  $\boldsymbol{\delta}_{k-1}$  utilizing  $f_{k-1}$  alone (cf. (12.26)). The quantity  $\hat{\boldsymbol{\delta}}_{k-1}$  is sometimes referred to as the smoothed value of  $\boldsymbol{\delta}_{k-1}$ ; the ‘present’ information  $\mathbf{l}_k$  is used to correct the value  $\boldsymbol{\delta}_{k-1}^{(1)}$  estimated in the past.

Let us return to the main problem, that of expressing the state vector  $\hat{\mathbf{x}}_k$  as a function of known quantities. To do this, (134) is used where  $\mathbf{x}_k^{(0)}$  is given by (140) and  $\hat{\boldsymbol{\delta}}_k$  has to be estimated from (144). Taking (144) and eliminating successively  $\hat{\boldsymbol{\delta}}_{k-1}$ ,  $\hat{\mathbf{k}}_{k-1, k}$ , and  $\hat{\mathbf{k}}_k$ , again using the technique shown in §3.1, one obtains

$$\hat{\boldsymbol{\delta}}_k = \boldsymbol{\delta}_k^{(1)} - \mathbf{G}_k (\mathbf{w}_k + \mathbf{A}_k \boldsymbol{\delta}_k^{(1)}), \quad (14.148)$$

where

$$\boldsymbol{\delta}_k^{(1)} = \mathbf{S}_{k-1, k} \hat{\boldsymbol{\delta}}_{k-1}. \quad (14.149)$$

The matrix  $\mathbf{G}_k$  has the special name of *gain matrix* and equals to

$$\mathbf{G}_k = \mathbf{C}_{x_k^{(0)}} \mathbf{A}_k^T (\mathbf{B}_k \mathbf{C}_{l_k} \mathbf{B}_k^T + \mathbf{A}_k \mathbf{C}_{x_k^{(0)}} \mathbf{A}_k^T)^{-1}, \quad (14.150)$$

where  $\mathbf{C}_{x_k^{(0)}}$  is the covariance matrix of the predicted state vector  $\mathbf{x}_k^{(0)}$  obtained simply by applying the covariance law to (140):

$$\mathbf{C}_{x_k^{(0)}} = \mathbf{C}_{\epsilon_{k-1, k}} + \mathbf{S}_{k-1, k} \mathbf{C}_{\hat{\mathbf{x}}_{k-1}} \mathbf{S}_{k-1, k}^T. \quad (14.151)$$

Finally, the covariance matrix of the estimated state vector  $\hat{\mathbf{x}}_k = \mathbf{x}_k^{(0)} + \hat{\boldsymbol{\delta}}_k$  is obtained by using the covariance law with (134). After extensive computations one obtains

$$\mathbf{C}_{\hat{\mathbf{x}}_k} = (\mathbf{I} - \mathbf{G}_k \mathbf{A}_k) \mathbf{C}_{x_k^{(0)}}. \quad (14.152)$$

Equations (140), (148), and (152) constitute the Kalman filter equations. They are applied in a recursive manner in this succession; the way the Kalman filter then works is shown in FIG. 13. It remains to be mentioned that the process of Kalman filtering starts with the selection of a convenient value for  $\mathbf{x}_0^{(0)}$  and the computation of  $\hat{\boldsymbol{\delta}}_0$ ,  $\hat{\mathbf{x}}_0$ , and  $\mathbf{C}_{\hat{\mathbf{x}}_0}$  from (131), (134), and (152). From there on the process proceeds in the way described above.

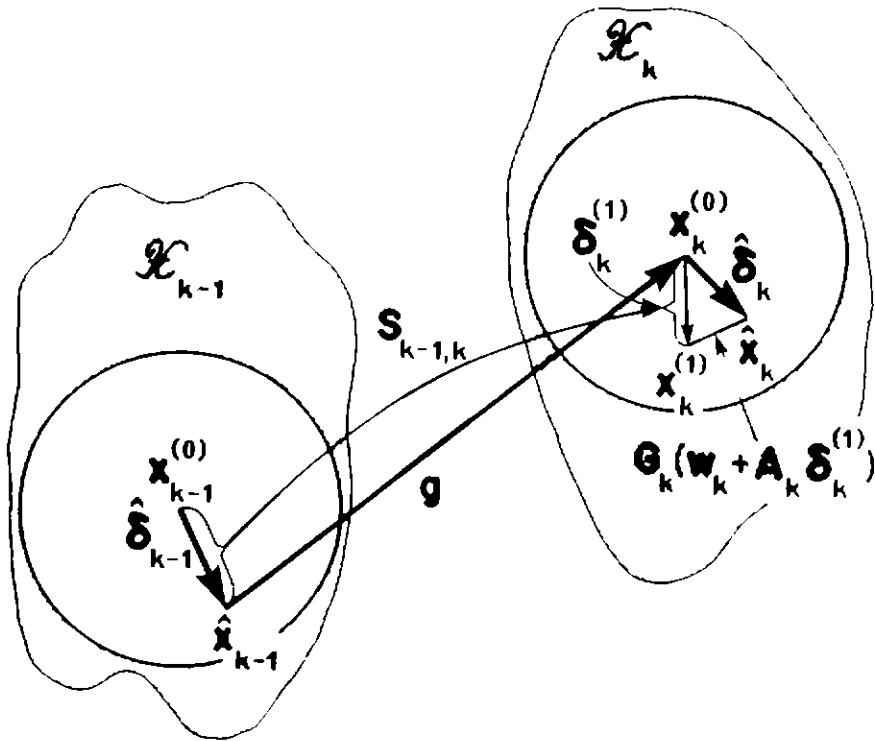


FIG. 14.13. One step of Kalman filter.

MORRISON [1969] gives equations of an alternative known as *Bayes filter* that, under certain circumstances, are more computationally efficient than the Kalman expressions. They are basically the same as those of Kalman except for the equation for the gain matrix, which reads

$$G_k = C_{\hat{x}_k} A_k^T M_k, \quad (14.153)$$

and the equation for the covariance matrix of the state vector,

$$C_{\hat{x}_k} = (C_{\hat{x}_{k-1}}^{-1} + N_k)^{-1}. \quad (14.154)$$

The Bayes and Kalman filtering equations are mathematically equivalent but are different from the computational point of view, each requiring a different number of operations. Their mathematical equivalence is easily shown by invoking the matrix lemmas ((3.22) and (3.23)). Application of the first to the Bayes expressions for the covariance matrix (154) results in an expression which corresponds exactly to the Kalman expression (152). The equivalence of the two gain matrices, (150) and (153), is shown by means of the second lemma.

As indicated at the outset of this section, the static procedures characterized by a fixed state vector (of unknown parameters) are simply special cases of the dynamic procedure. The *sequential equations* [SCHMID AND SCHMID, 1965] follow directly from the Kalman equations simply by realizing that the dynamic process (137) becomes a

static one by putting

$$\mathbf{x}_k^{(0)} = \hat{\mathbf{x}}_{k-1}, \quad (14.155)$$

so that  $\mathbf{S}_{k-1,k} = \mathbf{I}$ ,  $\mathbf{e}_{k-1,k} = \mathbf{0}$ , and  $\mathbf{C}_{\epsilon_{k-1,k}} = \mathbf{0}$ . One gets successively

$$\hat{\boldsymbol{\delta}}_k = \hat{\boldsymbol{\delta}}_{k-1} - \mathbf{G}_k (\mathbf{w}_k + \mathbf{A}_k \hat{\boldsymbol{\delta}}_{k-1}), \quad (14.156)$$

where

$$\mathbf{G}_k = \mathbf{C}_{\hat{\mathbf{x}}_{k-1}} \mathbf{A}_k^T (\mathbf{B}_k \mathbf{C}_{l_k} \mathbf{B}_k^T + \mathbf{A}_k \mathbf{C}_{\hat{\mathbf{x}}_{k-1}} \mathbf{A}_k^T)^{-1}, \quad (14.157)$$

and

$$\mathbf{C}_{\hat{\mathbf{x}}_k} = (\mathbf{I} - \mathbf{G}_k \mathbf{A}_k) \mathbf{C}_{\hat{\mathbf{x}}_{k-1}}. \quad (14.158)$$

Here,  $\hat{\boldsymbol{\delta}}_0$ ,  $\mathbf{C}_{\hat{\mathbf{x}}_0}$  are evaluated only from the first model (131). It has been shown by KRAKIWSKY [1975] that the above sequential equations are equivalent to those of TIENSTRA [1956]. Notice that in the above expressions (156) to (158) the effect of the 'new' data is given by an additional term.

Application of the same treatment to the Bayes filter equations yields the same expression (156) for  $\hat{\boldsymbol{\delta}}_k$ , where  $\mathbf{G}_k$  is given by (153), and  $\mathbf{C}_{\hat{\mathbf{x}}_k}$  is given by (154). Equations (156) and (154), together with the above equations for  $\hat{\boldsymbol{\delta}}_{k-1}$  and  $\mathbf{C}_{\hat{\mathbf{x}}_{k-1}}$ , constitute the *phase equations* approach.

It is instructive at this point to introduce a set of equations known as the *summation equations* which are mathematically equivalent to both the sequential and phase equations. They follow directly from (128) and (129): using the same technique as above, one gets

$$\hat{\boldsymbol{\delta}}_k = -(N_{k-1} + N_k)^{-1} (\mathbf{u}_{k-1} + \mathbf{u}_k) = -N^{-1} \mathbf{u}, \quad (14.159)$$

and

$$\mathbf{C}_{\hat{\boldsymbol{\delta}}_k} = (N_{k-1} + N_k)^{-1} = N^{-1}. \quad (14.160)$$

It is clear from inspecting the above equations why they are named the summation equations. Note that a matrix  $(N_{k-1}, N_{k-1} + N_k, \dots)$  of the order  $u$  must be inverted at each step of the process.

Since the Kalman and Bayes filters are mathematically equivalent for the treatment of a dynamic problem, and similarly the sequential, phase, and summation equations are equivalent for the treatment of a static case, one is tempted to ask whether it makes any difference to use one or the other process. It does from the computational point of view: the processes differ by the size of matrices to be inverted within each step. The summary of these differences (excluding the computations of  $\hat{\mathbf{x}}_0, \mathbf{C}_{\hat{\mathbf{x}}_0}$ ) is given in TABLE 3. It shows that the Kalman filter is more

**TABLE 14.3**  
 Sizes of matrices to be inverted in step-by-step procedures  
 $(m = \dim f, u = \dim x, n = \dim l)$

	Implicit or explicit models fully populated $C_{l_k}$	Explicit models diagonal $C_{l_k}$
Kalman	$m$	$m$
Bayes	$m, u$	$u$
Sequential	$m$	$m$
Phase	$m, u$	$u$
Summation	$n, u$	$u$

economical than the Bayes, and the sequential procedure is more economical than the other two approaches when statistically dependent observations are considered. The situation reverses when statistically independent observations and explicit models are used. When the matrices involved are sparse, the situation changes again; the interested reader is referred to KNIGHT AND MEPHAM [1978] for details.

## PART III

### REFERENCES

- ABRAMOWITZ, M. AND I.A. STEGUN (Eds.) (1964). *Handbook of Mathematical Functions*. Dover reprint, 1965.
- AHLBERG, J.H., E.N. NILSON AND J.L. WALSH (1967). *The Theory of Splines and their Applications*. Academic Press.
- BAARDA, W. (1967). Statistical concepts in geodesy. Netherlands Geodetic Commission, Publications on Geodesy, New Series 2 (4), Delft, Netherlands.
- BAARDA, W. (1976). Reliability and precision of networks. Delft Geodetic Institute, Publications of the Computing Centre, Delft, Netherlands.
- BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. Reprinted in *Biometrika* 45, pp. 293–315, 1958.
- BEN-ISRAEL, A. AND T.N.E. GREVILLE (1974). *Generalized Inverse: Theory and Applications*. Wiley-Interscience.
- BEREZIN, I.S. AND N.P. ZHIDKOV (1962). *Computing Methods*. Vol. I. Translated from Russian 2nd ed., 1965. Addison-Wesley.
- BJERHAMMAR, A. (1973). *Theory of Errors and Generalized Matrix Inverses*. Elsevier.
- BLACKMAN, R.B. AND J.W. TUKEY (1958). *The Measurement of Power Spectra from the Point of View of Communications Engineering*. Dover reprint, 1959.
- BLAHA, G. (1971). Inner adjustment constraints with emphasis on range observations. Department of Geodetic Science Report 148, The Ohio State University, Columbus, U.S.A.
- BLAHA, G. (1978). Personal communication.
- BLAHA, G. (1982). Notes on equivalent forms of the general least-squares solution. *Bull. Géod.* 56, pp. 220–230.
- BOMFORD, G. (1971). *Geodesy*. 3rd ed., Oxford University Press.
- BOSSLER, J.D. (1972). Bayesian inference in geodesy. Ph.D. dissertation, Department of Geodetic Science, The Ohio State University, Columbus, U.S.A. Revised, 1976.
- BOSSLER, J.D., E. GRAFARENDE AND R. KELM (1973). Optimal design of geodetic nets. 2. *J. Geophys. Res.* 78 (26), pp. 5887–5897.
- BURG, J.P. (1967). Maximum entropy spectral analysis. Paper presented at the 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City, U.S.A., October 31.
- BURNSIDE, C.D. (1971). *Electromagnetic Distance Measurement*. In series "Aspects of Modern Land Surveying", Ed. J.R. Smith, Crosby Lockwood.
- CAPON, J. (1969). High resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* 57 (8), pp. 1408–1418.
- CELMINS, A. (1973). Least squares adjustment with finite residuals for non-linear constraints and partially correlated data. U.S. Ballistic Research Laboratories Report 1658, Maryland, U.S.A.
- CHAUVENET, W. (1871). *A Manual of Spherical and Practical Astronomy: Theory and Use of Astronomical Instruments. Method of Least Squares*. Vol. II, 4th ed., Lippincott.
- CHENEY, E.W. (1966). *Introduction to Approximation Theory*. McGraw-Hill.
- COOK, A.H. (1973). *Physics of the Earth and Planets*. Macmillan.
- COOPER, M.A.R. (1974). *Fundamentals of Survey Measurement and Analysis*. Crosby Lockwood Staples.
- COTLAR, M. AND R. CIGNOLI (1974). *An Introduction to Functional Analysis*. Translation from Spanish by A. Torchinsky and A. Gonzalez Villalobos of 1974 ed. by Editorial Universitaria de Buenos Aires, North-Holland.

- CROSS, P.A. AND K. THAPA (1979). The optimal design of levelling networks. *Surv. Rev.* XXV (192), pp. 68-79.
- CROW, E.L., F.A. DAVIS AND M.W. MAXFIELD (1960). *Statistics Manual*. Dover.
- DAVIS, P.J. (1963). *Interpolation and Approximation*. Dover reprint, 1975.
- DIXON, W.J. (1962). Rejection of observations. In: *Contributions to Order Statistics*, Eds. A.E. Sarhan and B.G. Greenberg, Wiley.
- DRAPER, N.R. AND H. SMITH (1967). *Applied Regression Analysis*. Wiley.
- FALLER, J.E. (1965). An absolute interferometric determination of the acceleration of gravity. *Bull. Géod.* 77, pp. 203-204.
- FELLER, W. (1968). *An Introduction to Probability Theory and its Applications*. Vol. I, 3rd ed., Wiley.
- FORWARD, R.L. (1974). Review of artificial satellite gravity gradiometer techniques for geodesy. *Proc. International Symposium on the Use of Artificial Satellites for Geodesy and Geodynamics*, Ed. G. Veis. IAG and COSPAR, Athens, Greece, May 1973. National Technical University, pp. 157-192.
- GAUSS, K.F. (1809). *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*. Translation by Charles Henry Davis of Little Brown 1857 ed. Dover reprint, 1963.
- GODIN, G. (1972). *The Analysis of Tides*. University of Toronto Press.
- GOLD, B. AND C.M. RADAR (1969). *Digital Processing of Signals*. McGraw-Hill.
- GOLDMAN, S. (1953). *Information Theory*. Dover reprint, 1968.
- GRAFAREND, E.W. (1974). Optimization of geodetic networks. *Canad. Surv.* 28 (5), pp. 716-723.
- GRAFAREND, E. AND A. D'HONE (1978). Gewichtsschätzung in Geodätischen Netzen. Deutsche Geodätische Kommission, Reihe A, Heft Nr. 88, Munich, Germany.
- GRAFAREND, E. AND P. HARLAND (1973). Optimales Design geodätischer Netze, I. Deutsche Geodätische Kommission, Reihe A: Höhere Geodäsie, Heft Nr. 74, Munich, Germany.
- GRAYBILL, F.A. (1976). *Theory and Application of the Linear Model*. Duxbury.
- GUIER, W.H. AND G.C. WEIFFENBACH (1960). A satellite Doppler navigation system. *Proc. IRE* 48, April, pp. 507-516.
- HADLEY, G. (1964). *Nonlinear and Dynamic Programming*. Addison-Wesley.
- HAMILTON, W.C. (1967). *Statistics in Physical Science*. 2nd ed., Ronald.
- HANCOCK, H. (1917). *Theory of Maxima and Minima*. Dover reprint, 1960.
- HANSON, R.H. (1976). The new adjustment of the North American horizontal datum. *ACSM Bull.* 55, pp. 21-22.
- HIRVONEN, R.A. (1971). *Adjustment by Least Squares in Geodesy and Photogrammetry*. Ungar.
- HODGES, D.J. AND J.B. GREENWOOD (1971). *Optical Distance Measurement*. Butterworths.
- HOGG, R.V. AND A.T. CRAIG (1970). *Introduction to Mathematical Statistics*. 3rd ed., Macmillan.
- JEFFREYS, H. (1961). *Theory of Probability*. 3rd ed., Clarendon.
- JENKINS, G.M. AND D.G. WATTS (1968). *Spectral Analysis and its Applications*. Holden-Day.
- KALMAN, R.E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Engrg.-Trans. ASME*, March, pp. 35-45.
- KNIGHT, W. AND M.P. MEPHAM (1978). Report on computer programs for solving large systems of normal equations: *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 357-363.
- KNIGHT, W. AND P. STEEVES (1974). Partial solution of the variance-covariance matrix of geodetic networks. *Canad. Surv.* 28 (5), pp. 686-689.
- KORN, G.A. AND T.M. KORN (1968). *Mathematical Handbook for Scientists and Engineers*. 2nd ed., McGraw-Hill.
- KRAKIWSKY, E.J. (1975). A synthesis of recent advances in the method of least squares. Department of Surveying Engineering Lecture Note 42, University of New Brunswick, Fredericton, Canada.
- KRARUP, T. (1969). A contribution to the mathematical foundation of physical geodesy. Danish Geodetic Institute Publication No. 44, Copenhagen, Denmark.
- LANCZOS, C. (1957). *Applied Analysis*. Pitman.
- LAWSON, C.L. AND R.J. HANSON (1974). *Solving Least Squares Problems*. Prentice-Hall.
- LEHR, C.G., C.R.H. TSIANG, G.M. MENDES AND R.J. ELDRED (1974). Laser pulse analysis. *Proc.*

- International Symposium on the Use of Artificial Satellites for Geodesy and Geodynamics*, Ed. G. Veis, IAG and COSPAR, Athens, Greece, May 1973. National Technical University; pp. 109–118.
- LENNON, G.W. (1970). Sea level instrumentation, its limitations and the optimisation of the performance of conventional gauges in Great Britain. *Report on the Symposium on Coastal Geodesy*, Ed. R. Sigl. IUGG, IAG, Munich, Germany, July. Institut für Angewandte Geodäsie, pp. 181–200.
- LENNON, G.W. (1974). Mean sea level as a reference for geodetic leveling. *Canad. Surv.* 28 (5), pp. 524–530.
- LIEBELT, P.B. (1967). *An Introduction to Optimal Estimation*. Addison-Wesley.
- LUENBERGER, D.G. (1969). *Optimization by Vector Space Methods*. Wiley.
- MAGNESS, T.A. AND J.B. MCGUIRE (1962). Comparison of least squares and minimum variance estimates of regression parameters. *Ann. Math. Statist.* 33 (2), pp. 462–470.
- MELCHIOR, P. (1978). *The Tides of the Planet Earth*. Pergamon.
- MIKHAIL, E.M. (1976). *Observations and Least Squares*. IEP-A dun-Donnelley Publisher.
- MILLER, R.G. (1966). *Simultaneous Statistical Inference*. McGraw-Hill.
- MORITZ, H. (1972). Advanced least squares methods. Department of Geodetic Science Report 175, The Ohio State University, Columbus, U.S.A.
- MORITZ, H. (1973). Stepwise and sequential collocation. Department of Geodetic Science Report 203, The Ohio State University, Columbus, U.S.A.
- MORRISON, N. (1969). *Introduction to Sequential Smoothing and Prediction*. McGraw-Hill.
- MUELLER, I.I. (1963). Geodesy and the torsion balance. *J. Surv. Map. Div. Proc. Am. Soc. Civ. Engrg.* 89, pp. 123–155.
- MUELLER, I.I. (1964). *Introduction to Satellite Geodesy*. Ungar.
- MUELLER, I.I. (1969). *Spherical and Practical Astronomy as Applied to Geodesy*. Ungar.
- NICKERSON, B.G., E.J. KRAKIWSKY, D.B. THOMSON, M.L. SYVVERSON-KRAKIWSKY AND J.M. CRAWFORD (1978). Design of survey networks using interactive computer graphics. *Proc. 38th Annual Meeting of the American Congress on Surveying and Mapping*, Washington, D.C., U.S.A., February, pp. 378–388.
- OTNES, R.K. AND L. ENOCHSON (1972). *Digital Time Series Analysis*. Wiley.
- PARTHASARATHY, K.R. AND K. SCHMIDT (1972). *Positive Definite Kernels, Continuous Tensor Products, and Central Limit Theorems of Probability Theory*. In series “Lecture Notes in Mathematics”, Eds. A. Dold and B. Eckmann, Springer.
- PERELMUTER, A. (1979). Adjustment of free networks. *Bull. Géod.* 53 (4), pp. 291–295.
- POPE, A.J. (1974). Two approaches to nonlinear least squares adjustments. *Canad. Surv.* 28 (5), pp. 663–669.
- POPE, A.J. (1976). The statistics of residuals and the detection of outliers. NOAA Technical Report NOS 65 NGS 1, U.S. Department of Commerce, Rockville, U.S.A.
- QUESENBERRY, C. AND H. DAVID (1961). Some tests for outliers. *Biometrika*, 48, pp. 379–390.
- RAO, C.R. AND S.K. MITRA (1971). *Generalized Inverse of Matrices and its Applications*. Wiley.
- REVUZ, D. (1975). *Markov Chains*. North-Holland.
- ROBBINS, A.R. (1976). Military engineering: Field and geodetic astronomy. Vol. 13, Part 9, Ministry of Defence Army Code No. 71091, School of Military Survey, Hermitage, Newbury, Berkshire, U.K.
- SAVAGE, I.R. (1953). Bibliography of nonparametric statistics and related topics. *J. Am. Statist. Assoc.* 48, pp. 844–906.
- SCHMID, H.H. AND E. SCHMID (1965). A generalized least squares solution for hybrid measuring systems. *Canad. Surv.* 19 (1), pp. 27–41.
- SCHWARZ, C.R. (1969). The use of short arc orbital constraints in the adjustment of geodetic satellite data. Department of Geodetic Science Report 118, The Ohio State University, Columbus, U.S.A.
- SIEGEL, S. (1956). *Nonparametric Statistics: For the Behavioral Sciences*. McGraw-Hill.
- SINGER, I. (1970). *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*. Translated from Russian by R. Georgescu, Springer.
- SMITH, J.R. (1970). *Optical Distance Measurement*. In series “Aspects of Modern Land Surveying”, Ed. J.R. Smith, Crosby Lockwood.
- STEEVES, R.R. (1978). A note on the optimal design of geodetic networks. Personal communication, Department of Surveying Engineering, University of New Brunswick, Fredericton, Canada.
- STEFANSKY, W. (1972). Rejecting outliers in factorial designs. *Technometrics* 14 (2), pp. 469–479.

- SYNGE, J.L. AND A. SCHILD (1949). *Tensor Calculus*. University of Toronto Press.
- THEIL, H. (1963). On the use of incomplete prior information in regression analysis. *J. Am. Statist. Assoc.* 58, pp. 401–414.
- THOMPSON, E.H. (1969). *An Introduction to the Algebra of Matrices with some Applications*. University of Toronto Press.
- THOMPSON, W. (1935). On the criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *Ann. Math. Statist.* 6, pp. 214–219.
- TIENSTRA, J.M. (1956). *Theory of the Adjustment of Normally Distributed Observations*. Edited by his friends. N.V. Uitgeverij Argus.
- VALI, V., R.S. KROGSTAD AND R.W. MOSS (1965). Laser interferometer for earth strain measurement. *Rev. Sci. Instrum.* 36, pp. 1352–1355.
- VANÍČEK, P. (1971). Further development and properties of the spectral analysis by least-squares. *Astrophys. and Space Sci.* 12, pp. 10–33.
- VEIS, G. (ED.) (1963). *The Use of Artificial Satellites for Geodesy*. Proceedings of the First International Symposium on the Use of Artificial Satellites for Geodesy. COSPAR IUGG, Washington, D.C., U.S.A., April, 1962. North-Holland.
- WILKS, S.S. (1962). *Mathematical Statistics*. Wiley.
- WILLKE, T.A. (1965). Useful alternatives to Chauvenet's rule for rejection of measurement data. Statistical Engineering Laboratory Report, U.S. National Bureau of Standards, Washington, D.C., U.S.A.
- WONNACOTT, T.H. AND R.J. WONNACOTT (1972). *Introductory Statistics*. 2nd ed., Wiley.
- WREDE, R.C. (1963). *Introduction to Vector and Tensor Analysis*. Dover reprint, 1972.
- ZYGMUND, A. (1968). *Trigonometric Series*. Vol. 1 and 2, Cambridge University Press.

**PART IV**

**POSITIONING**

## CHAPTER 15

### POINT POSITIONING

The determination of the coordinates of a point on land, at sea, or in space with respect to an implied coordinate system is called point positioning. The problem of point positioning may be stated as follows: Given the coordinates of observed extraterrestrial objects, such as stars or satellites, along with the measurements of quantities linking a terrestrial point to these objects, compute the coordinates of the point. Positioning of a point with respect to other terrestrial points is treated in Chapter 16.

Because point positioning can be done in three different modes, three distinctly different classes of coordinate systems are needed: terrestrial coordinate systems for earth-located points to be positioned; celestial coordinate systems for the sighted stars; and orbital coordinate systems for the observed satellites. Fundamental to the definitions of these coordinate systems are the motions of the earth and satellites in space. The earth itself has two main periodic motions of importance to us here: it revolves about the sun, and it spins about its own axis (cf. Chapter 5). Its only natural satellite (the moon) and many artificial satellites move independently about the earth. Terrestrial coordinate systems are earth-fixed: they both spin and revolve with the earth. Celestial coordinate systems do not revolve but may spin with the same velocity as the earth. The orbital coordinate systems do not spin with the earth but revolve with it.

In the current chapter, the first section contains the fundamentals of astronomical positioning consisting mainly of the definitions of the principal celestial coordinate systems. The second section treats the mathematical models used in the astronomical determination of coordinates. As well, the models for astronomical azimuths are shown. The third section deals with the mathematical models for determining the position of a point from observations to satellites, i.e., with satellite positioning. The fourth section addresses the idea of positioning of the reference ellipsoid. Transformation of positions from one ellipsoid onto another are also discussed, as well as the mapping of the ellipsoid onto a plane. In the models shown here, no allowance for time deformations of the earth is made. Throughout this chapter, the point to be positioned is considered to be stationary with respect to the earth; positioning of a moving point, as part of navigation, is discussed in Chapter 16.

### 15.1. Fundamentals of geodetic astronomy

Let us begin by discussing the celestial systems of coordinates. Since the distance from the earth to the nearest star (excluding the sun) is more than  $10^9$  larger than the earth's radius, the dimension of the earth is negligible compared with the distance to stars. The stars in our galaxy are almost immobile, however the galaxies themselves are believed to be moving at velocities comparable to the velocity of light. To an observer on the earth, though, even this motion is perceived to be very slow causing a displacement that rarely exceeds one second of arc per year. Therefore, one may consider the stars and galaxies to be located on a surface called the *celestial sphere* (see §1.1) the dimension of which is so large that the earth can be considered dimensionless (as a point) at the centre of this sphere, much the same way as it was in §5.1. Directions on the earth and in the solar system can then be extended to the celestial sphere. Points and curves on the celestial sphere obtained in this manner form the basis for the definition of all celestial coordinate systems.

The earth's (precessing and nutating) instantaneous spin axis (see §5.2) is extended outward to intersect the celestial sphere at the *north celestial pole* (NCP) and *south celestial pole* (SCP)—see FIG. 1. The earth's equatorial plane (the plane perpendicular to the spin axis and containing the centre of mass *C* of the earth)

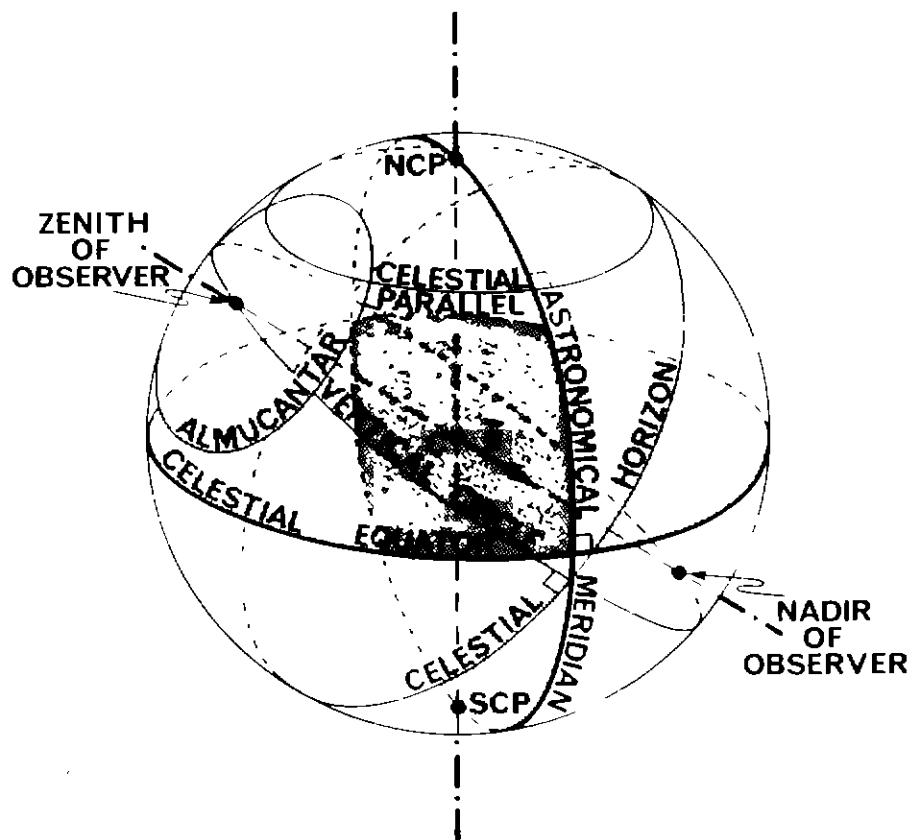


FIG. 15.1. Celestial sphere.

intersects the celestial sphere to form the *celestial equator*. A plane parallel to the celestial equator intersects the celestial sphere in a small circle called a *celestial parallel*. Any great circle containing the poles is perpendicular to both the celestial equator and parallels and is called an *astronomical (celestial) meridian*. The gravity vector  $\bar{g}$  of the observer extended upward intersects the celestial sphere at a point called the *zenith* (of the observer), and downward to define a point called the *nadir* (of the observer). The great circle made by a plane perpendicular to the observer's gravity vector is the *celestial horizon*. A small circle made by a plane parallel to the celestial horizon is called an *almucantar*. Any plane that contains the gravity vector is a vertical plane whose intersection with the celestial sphere is the *vertical circle*. The vertical plane normal to the astronomical meridian is called the *prime vertical* (or decuman); it intersects the celestial horizon to define the *east* and *west directions*.

To define the position of a star anywhere on the celestial sphere, all one really needs to know is a direction. A direction is most simply defined as a unit vector in polar coordinates ( $r, \theta, \lambda$ —see §3.3): since the first coordinate  $r$  always equals one, in effect the vector is specified by only the two angles. This is the approach adopted here. All the celestial systems will be considered spherical, defined by the location of the origin and directions of the axes of the representative Cartesian systems.

Let us begin with the *right ascension system* (RA), the most important of the *celestial* systems (see §1.1). It is *heliocentric*, i.e., its origin is the sun ( $H$ ), the  $z^{\text{RA}}$ -axis is directed toward the NCP, the  $x^{\text{RA}}$ -axis points toward the vernal point  $\gamma$  (see §5.2), and the system is right-handed—cf. FIG. 2. The *declination*  $\delta$  of the star ( $S$ ) is the angle between the celestial equatorial plane and the direction from  $H$  to  $S$ ,

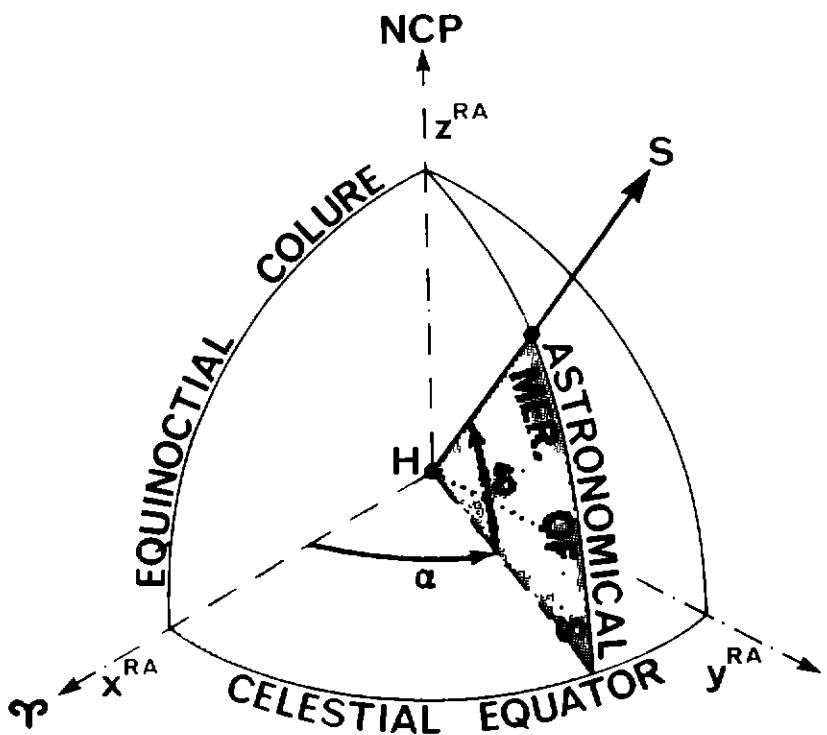


FIG. 15.2. Right ascension system.

measured in the astronomical meridian plane of  $S$ . The *right ascension*  $\alpha$  of  $S$  is the angle measured counterclockwise, as seen from the NCP, in the equatorial plane from  $\gamma$  to the astronomical meridian of  $S$ . Note that the astronomical meridian of  $\gamma$  is called the *equinoctial colure*. The unit vector describing the direction to  $S$  in this system is

$$\hat{e}^{\text{RA}} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}^{\text{RA}} = \begin{bmatrix} \cos \delta \cos \alpha \\ \cos \delta \sin \alpha \\ \sin \delta \end{bmatrix}, \quad (15.1)$$

while the angles are related to the Cartesian components by

$$\begin{aligned} \delta &= \arcsin z^{\text{RA}}, \\ \alpha &= \arctan(y^{\text{RA}}/x^{\text{RA}}). \end{aligned} \quad (15.2)$$

Since  $\alpha \in (0, 2\pi)$ , the second equation carries with it an uncertainty of  $\pi$ . It is thus sometimes preferable to use the equivalent equation for a half-angle,

$$\alpha = 2 \arctan \frac{y^{\text{RA}}}{x^{\text{RA}} + \sqrt{(x^{\text{RA}})^2 + (y^{\text{RA}})^2}}, \quad (15.3)$$

which is unequivocal.

It is in this system that positions of stars are published. There are, however, some complications involved here: clearly, because of precession and nutation (cf. §5.2), the NCP, being defined through the earth's instantaneous spin axis, moves among the stars as a function of time. Thus, the coordinate system changes with time as do the coordinates  $(\alpha, \delta)$  of the stars. In publishing the stars' positions, it is therefore necessary to specify the epoch  $\tau_0$  to which the coordinates refer. It is usual for star catalogues to use an RA system that precesses but does not nutate. This RA system is called a *mean right ascension system* — MRA( $\tau_0$ ).

The next most important coordinate system is the one in which the observations to stars are made. This system is defined through the observer's gravity vector and the direction of the earth's (conventional) spin axis. These two directions can be sensed by various astronomical instruments, and the *vertical (altitude) angle*  $\nu$ , *zenith distance*  $Z$ , and *astronomical azimuth*  $A$  can be directly measured (see FIG. 3). The gravity vector defines the negative  $z^{\text{LA}}$ -axis, and together with a parallel to the conventional spin axis (cf. §5.4) they define the  $xz^{\text{LA}}$ -plane of the system. The  $y^{\text{LA}}$ -axis completes the left-handed system. It is called the *local astronomical system* (LA), and its origin is at the site of the observer on the surface of the earth ( $T$ ); it is

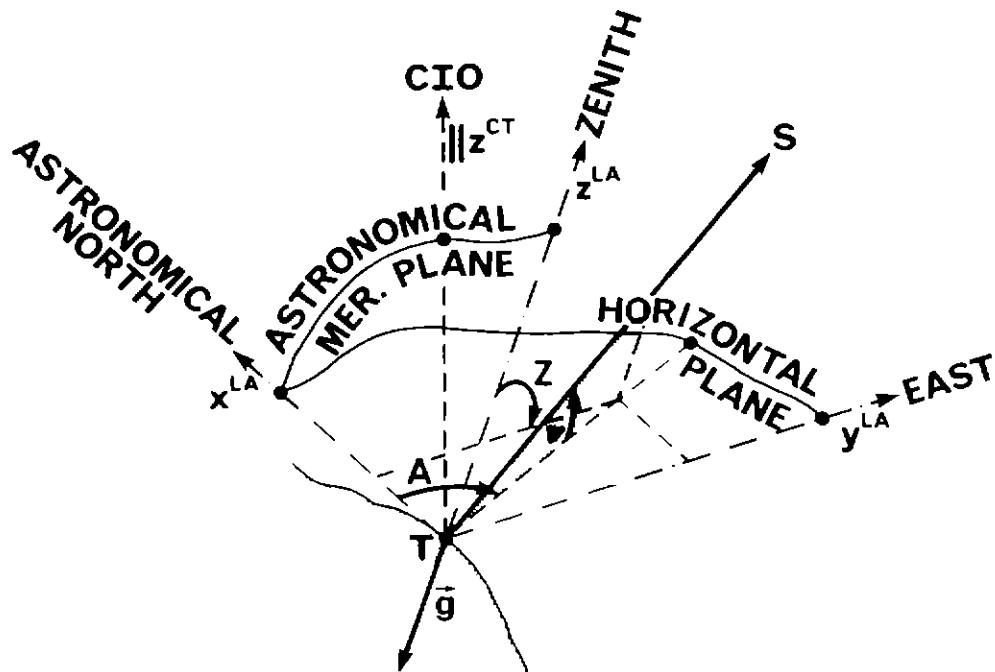


FIG. 15.3. Local astronomical system.

thus said to be *topocentric*. In this system, the unit vector in the direction of  $S$  is

$$\bar{e}^{LA} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}^{LA} = \begin{bmatrix} \cos \nu \cos A \\ \cos \nu \sin A \\ \sin \nu \end{bmatrix}, \quad (15.4)$$

while the angles are related to the Cartesian coordinates through

$$\nu = \frac{1}{2}\pi - Z = \arcsin z^{LA}, \quad A = 2 \arctan \frac{y^{LA}}{x^{LA} + \sqrt{(x^{LA})^2 + (y^{LA})^2}}. \quad (15.5)$$

Note that the system is not defined for those points whose gravity vector direction coincides with the direction of the conventional spin axis.

The observations are made in the topocentric LA system which is spinning as well as revolving with the earth. On the other hand, the star positions are usually given in the MRA( $\tau_0$ ) system that refers to an epoch (different from that of the observations) and which is motionless—up to the precession. The problem faced in the astronomical determination of positions is in the reconciliation of the above two systems. This problem is normally solved through a series of transformations from one system to

another. More accurately, the observations and the star coordinates are both transformed into a third coordinate system—the system of apparent places—to be defined later. The position of the observer on the earth's surface is then determined as a by-product of these transformations. Let us now go through these transformations, step by step ((a) to (f), see FIG. 12), starting with the LA system.

(a) The first coordinate system needed in transforming an LA system into the system of apparent places is the *conventional terrestrial system* (CT). The CT system is the closest practical approximation of the geocentric natural system described in §5.3 and is probably the most important system in geodesy. Its origin is at the centre of mass of the earth, the  $z^{\text{CT}}$ -axis points to the CIO (see §5.4), the  $xz^{\text{CT}}$ -plane contains the *mean Greenwich Observatory* [ROBBINS, 1976], and the  $y^{\text{CT}}$ -axis is selected to make the system right-handed (cf. FIG. 4).

The unit vector in the direction of the local zenith is again given by formulae similar to those for the RA and LA systems, which involve *astronomical latitude*  $\Phi$  and *astronomical longitude*  $\Lambda$  (see FIG. 4). These two angles are the ones used to define the position of a point astronomically. The principles of transformation (§3.3) can now be applied to obtain the 'observed' unit vector in the direction of  $S$  (eqn.(4))

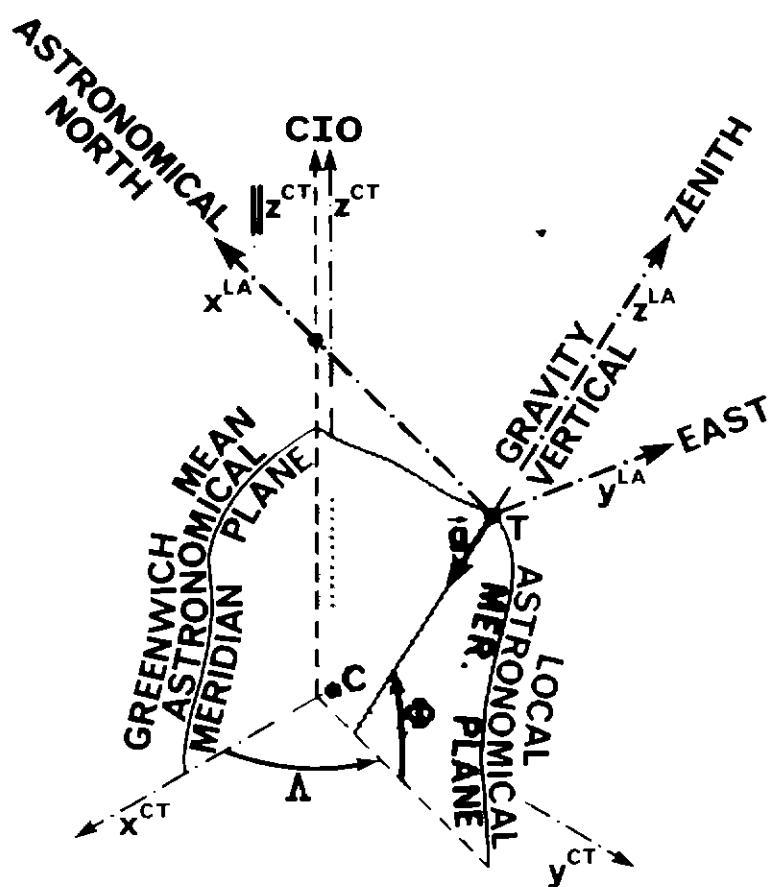


FIG. 15.4. Conventional terrestrial system.

in the CT system. We get

$$\begin{aligned}\bar{e}^{CT} &= \mathbf{R}_3(\pi - \Lambda) \mathbf{R}_2\left(\frac{1}{2}\pi - \Phi\right) \mathbf{P}_2 \bar{e}^{LA} \\ &= \begin{bmatrix} -\sin \Phi \cos \Lambda & -\sin \Lambda & \cos \Phi \cos \Lambda \\ -\sin \Phi \sin \Lambda & \cos \Lambda & \cos \Phi \sin \Lambda \\ \cos \Phi & 0 & \sin \Phi \end{bmatrix} \bar{e}^{LA}. \quad (15.6)\end{aligned}$$

It should be noted that the astronomical meridian plane of the observer contains both the gravity vector of the observer and the CIO; it is thus parallel to the conventional spin axis but generally does not contain the centre of mass of the earth.

(b) Next, the CT system gets transformed into the *instantaneous terrestrial system* (IT) that differs from the CT system only in so far as its  $z^{IT}$ -axis coincides with the instantaneous rather than the conventional spin axis. Thus, the only difference between these two systems is that the  $z^{IT}$ -axis wobbles around the  $z^{CT}$ -axis, and this wobble is described by the two parameters  $x_p, y_p$  in angular units (cf. §5.4), this situation is shown in FIG. 5. The transformation is effected through the following expression:

$$\bar{e}^{IT} = \mathbf{R}_1(y_p) \mathbf{R}_2(x_p) \bar{e}^{CT}. \quad (15.7)$$

Developing the trigonometrical functions into power series and neglecting second and higher order terms (as shown already,  $x_p, y_p$  are of the order of a few tenths of

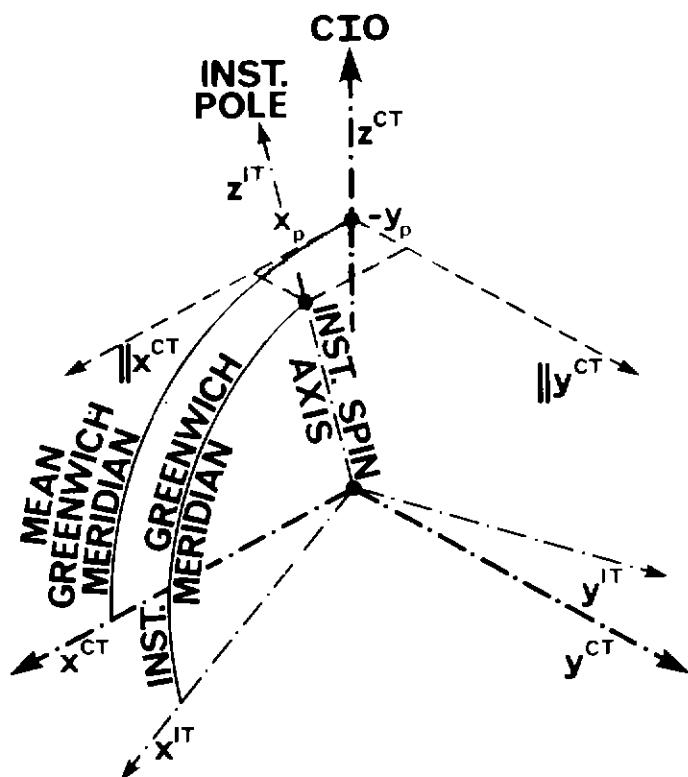


FIG. 15.5. Conventional and instantaneous terrestrial systems.

second of arc), we get

$$\bar{e}^{IT} = \begin{bmatrix} 1 & 0 & -x_P \\ 0 & 1 & y_P \\ x_P & -y_P & 1 \end{bmatrix} \bar{e}^{CT}. \quad (15.8)$$

Clearly, the IT system changes its position within the earth with time—even its  $x^{IT}$ -axis moves with time to allow the  $xz^{IT}$  plane to pass through the '*instantaneous Greenwich Observatory*' [ROBBINS, 1976]—and its epoch should be that of the observations made. It should be equally clear that if the observed azimuth  $A$  happens to be the *instantaneous astronomical azimuth*  $A(\tau)$ —which refers to the instantaneous spin axis instead of to the direction to CIO—then the observations made in the LA system should be transformed directly to the  $IT(\tau)$  system using (6) where  $\bar{e}^{CT}$  is replaced by  $\bar{e}^{IT}$ . As will be shown later, there are measuring systems that provide us directly with  $A(\tau)$ .

(c) The last transformation is from the  $IT(\tau)$  system into the *apparent place system* ( $AP(\tau)$ ). The  $AP(\tau)$  system is another geocentric system in which the  $z^{AP}$ -axis coincides with the  $z^{IT}$ -axis, and  $x^{AP}$ -axis points toward  $\varphi$ , and  $y^{AP}$ -axis completes the system to make it right-handed. The situation is shown in FIG. 6. Clearly, the transformation from the  $IT(\tau)$  to the  $AP(\tau)$  system consists of rotating the IT system around the common  $z$ -axis by the angle known as *Greenwich apparent sidereal time* (GAST), i.e.,

$$\bar{e}^{AP} = R_3(-GAST) \bar{e}^{IT}. \quad (15.9)$$

The other symbols appearing on this figure will be explained a little later.

(d) Let us now return to the  $MRA(\tau_0)$  system and again follow the path of transformations leading to the  $AP(\tau)$  system but from the other side (cf. FIG. 12).

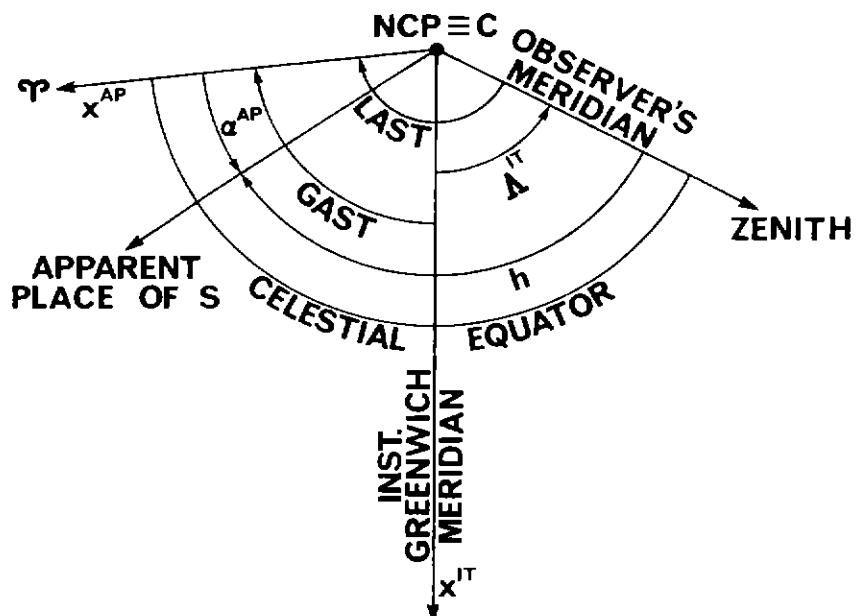


FIG. 15.6. Sidereal time, hour angle, right ascension, and longitude.

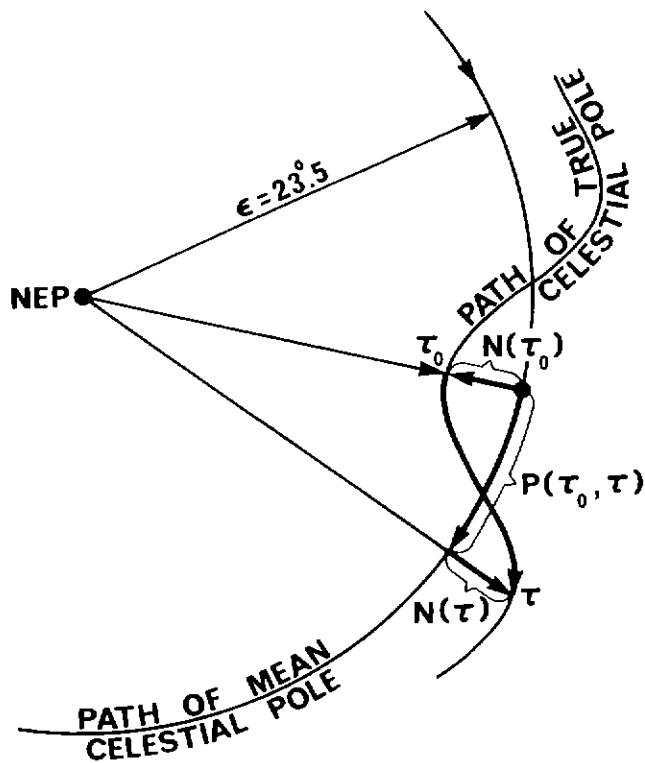


FIG. 15.7. Path of the celestial pole.

The first step is to transform the MRA system from epoch  $\tau_0$ , in which the catalogue is published, to  $\tau$ , when the observations are made. To update the coordinates  $\alpha^{\text{MRA}}(\tau_0)$ ,  $\delta^{\text{MRA}}(\tau_0)$  to epoch  $\tau$ , one must account for two effects: the effect of precession during the time  $\tau - \tau_0$  (see FIG. 7), and the effect of proper motion of the stars during the same period.

The amount of precession  $P(\tau_0, \tau)$  occurring in a time interval  $(\tau_0, \tau)$  is usually spelled out in terms of three *precessional constants*  $(\zeta_0, \theta, z)$  as shown in FIG. 8. Expressions for these elements as functions of time were derived early in this century by NEWCOMB [1906]. The angles  $(\frac{1}{2}\pi - \zeta_0)$  and  $(\frac{1}{2}\pi + z)$  are the right ascensions of the ascending node of the mean equator at  $\tau$ , measured respectively in the two mean systems (at  $\tau_0$  and  $\tau$ ). The angle  $\theta$  is the inclination between the mean equators at  $\tau$  and at  $\tau_0$ . The transformation of  $\alpha, \delta$  from  $\tau_0$  to  $\tau$  is made first by transforming  $\alpha$  and  $\delta$  into Cartesian components of the corresponding unit vector (1), then rotating this vector as follows:

$$\bar{e}^{\text{MRA}(\tau)} = \mathbf{R}_3(-z) \mathbf{R}_2(\theta) \mathbf{R}_3(-\zeta_0) \bar{e}^{\text{MRA}(\tau_0)}, \quad (15.10)$$

and then back into  $\alpha$  and  $\delta$  by means of (2).

In addition to the apparent motion of the MRA system due to the precession, the stars have a motion of their own called *proper motion*. Because this motion appears, for all practical purposes, to be linear, it is most appropriate to account for it in an almost linearly moving system. Accordingly, proper motion, usually tabulated as constant rates of change in right ascension and declination for each star of interest, is taken care of within the transformation  $\text{MRA}(\tau_0) \rightarrow \text{MRA}(\tau)$  [MUELLER, 1969].

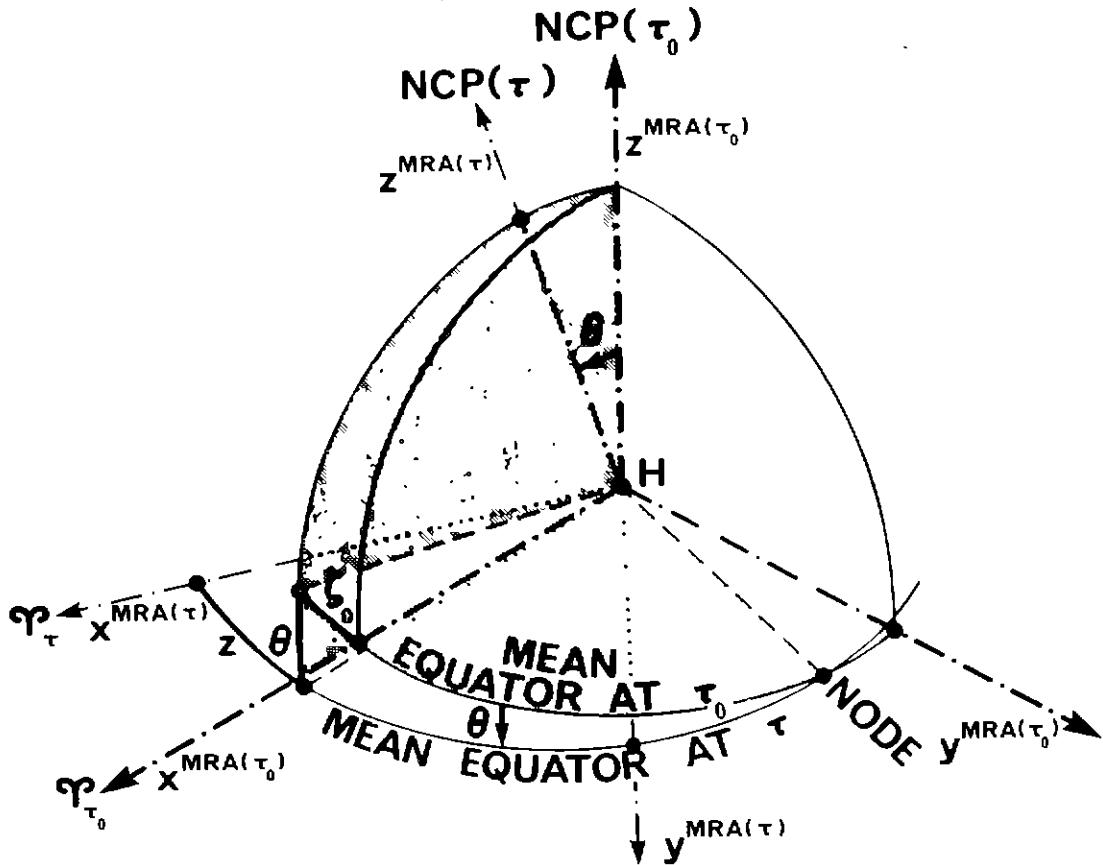


FIG. 15.8. Mean right ascension system and precession.

(e) The next step in updating the star coordinates is to account for the nutation  $N(\tau)$  (FIG. 7). This step defines the *true right ascension system* at epoch  $\tau$ —TRA( $\tau$ )—whose  $z^{\text{TRA}}$ -axis coincides with the instantaneous spin axis of the earth, while the true vernal equinox defines the direction of the  $x^{\text{TRA}}$ -axis. The effect of nutation  $N(\tau)$  is usually spelled out in terms of *nutation in longitude*  $\Delta\psi$  and *nutation in the obliquity*  $\Delta\epsilon$  (FIG. 9). The transformation of  $\alpha$  and  $\delta$  from the MRA( $\tau$ ) to the TRA( $\tau$ ) system is accomplished by first rotating the Cartesian components of the appropriate unit vector, i.e.,

$$\bar{e}^{\text{TRA}(\tau)} = \mathbf{R}_1(-\epsilon - \Delta\epsilon) \mathbf{R}_3(-\Delta\psi) \mathbf{R}_1(\epsilon) \bar{e}^{\text{MRA}(\tau)}, \quad (15.11)$$

and then transforming these back into  $\alpha$  and  $\delta$  by means of (2). The obliquity angle  $\epsilon$  has already been defined in §5.2.

(f) The last step in the chain of updates (transformations) taking us to the AP( $\tau$ ) system once more is to account for the fact that the star is not observed from the origin  $H$  (i.e., from the centre of the sun) of the RA system, but from the earth. The heliocentric values of  $\alpha$  and  $\delta$  must receive a correction, called *annual parallax*, that can be expressed as the parallactic angle of the radius of the earth's orbit subtended at the star. The parallactic corrections to  $\alpha$  and  $\delta$  are obtained from simple expressions that reflect the position and distance of the star; for the nearest star, the correction is about  $0.8''$  [MUELLER, 1969].

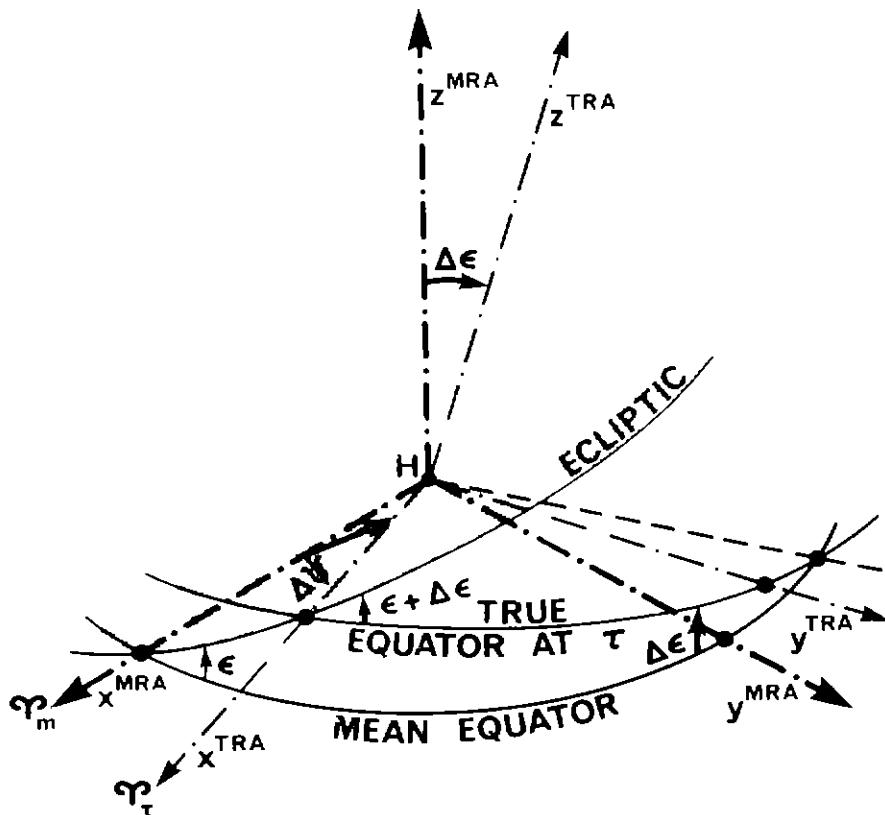


FIG. 15.9. True and mean right ascension systems.

Further, due to the fact that the observations are made from an orbiting (moving) earth, the light from a star will appear to be coming from a slightly different direction than it actually does—see FIG. 10. This effect is called *annual aberration* and does not exceed a value of  $20''$ ; it is calculated by means of the constant of (annual) aberration,  $v/c$ , where  $v$  stands for the velocity of the earth in its orbit, and  $c$  is the velocity of light. In most cases, the above value is only a few seconds of arc as a result of the relative configuration of the earth, sun, and the star in question [SMART, 1962]. It should be mentioned here that all the above updates become unnecessary when a special star catalogue, called *Apparent Places of Fundamental Stars* (APFS) (see, e.g., APFS, 1979 [1977]), is used instead of the standard fundamental catalogues.

Now let us return to the first step (a) of the transformation chain  $\text{LA}(\tau) \rightarrow \text{AP}(\tau)$  (see FIG. 12) and consider what has to be done to the observations to make them compatible with the apparent places of the stars. There are three effects to be considered in this context: diurnal aberration, diurnal parallax, and refraction. *Diurnal aberration* is a result, again, of making measurements from an earthbound observing station spinning with the earth. It is the translation velocity  $v$  of the station that causes the star to undergo an apparent shift analogous to the one shown in FIG. 10. The constant of diurnal aberration  $v/c$  ( $v < 2\pi R/\text{day} = 463 \text{ ms}^{-1}$ ) corresponds to a maximum correction of  $0.3''$  on the equator. It is rather small, but as it is systematic it should be removed from the observations. *Diurnal parallax* is caused by the parallactic angle of the earth's radius subtended at the star, and it is

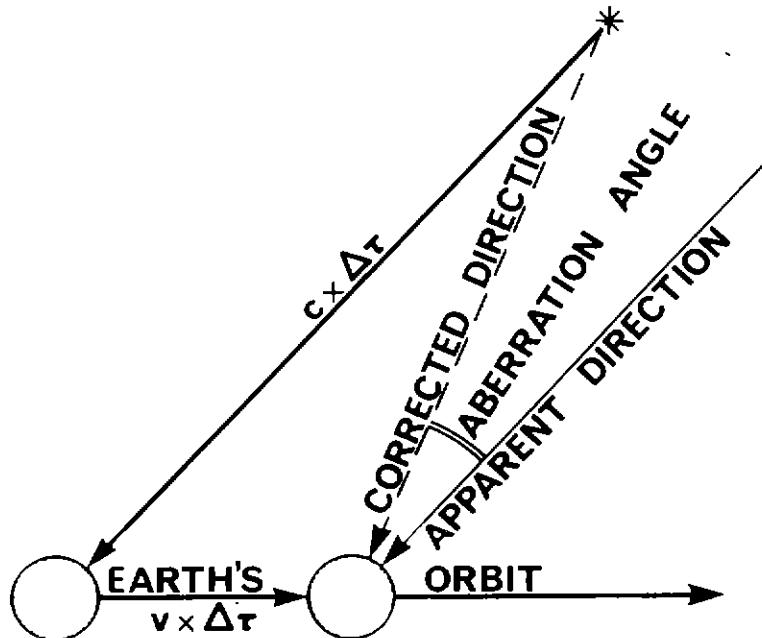


FIG. 15.10. Aberration.

always negligibly small. *Astronomical refraction*, the bending of the light rays from the stars as they enter the earth's atmosphere (cf. §9.2), seriously affects zenith distance measurements and, to a lesser degree, even the azimuth measurements. This effect will be discussed in the context of the mathematical models developed in §15.2.

Sometimes there is a need for a coordinate system that is motionless with respect to the galaxies. Such a system is called an *inertial system* and has the property of generating no acceleration on objects reckoned in that system. A good approximation of such an inertial system is the *ecliptical system* (E). It is heliocentric, its  $z^E$ -axis coincides with the earth's precession axis (see §5.2), the  $x^E$ -axis points toward  $\Upsilon$ , and  $y^E$  is chosen to make the system right handed (see FIG. 11). It uses *ecliptical latitude*  $\beta$  and *ecliptical longitude*  $\lambda^E$  as shown. This system is almost inertial except for the precessional shift of (see §5.2), the *planetary precession*, i.e., the precession caused by the planets in the solar system, and the very slow movement with the galaxy (see §5.1); it would be the best one to use for star coordinate publications. However, the difference between the inertialness of the MRA and E systems is insignificant in practice, and the E system is normally not used.

The last concept essential in astronomical positioning is the concept of time. Odd as it may seem, time in astronomy may be interpreted as merely an angle between two corresponding axes of two particular coordinate systems, as we have already seen in FIG. 6. in the case of the GAST. The *hour angle*  $h$  of  $S$  is the angle between the astronomical meridian of  $S$  and that of the observer. The *local apparent sidereal time* (LAST) is the hour angle of the true vernal equinox; similarly, the GAST is the hour angle of the true vernal equinox, as seen at Greenwich. The GAST and LAST are linked together by the expression

$$\text{LAST} = \text{GAST} + \Lambda^{\text{IT}}. \quad (15.12)$$

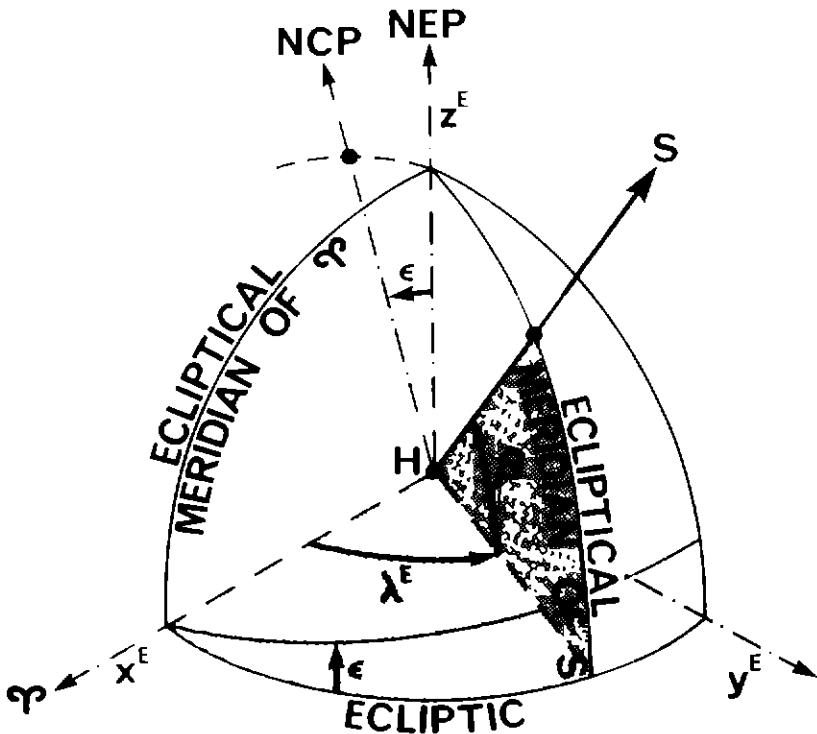


FIG. 15.11. Ecliptical system.

It is interesting to note that the  $x^{\text{TRA}}$ -axis is fixed in space (up to precession and nutation), while the  $x^{\text{IT}}$ -axis is attached to the earth and spins with it at the sidereal rate, i.e., the actual irregular rotation rate of the earth (cf. §5.4).

In practice, the GAST is measured through the *universal time* (UT) which in turn differs from every day *standard time* by an integral number of hours depending on one's time zone. Several different UTs are used [MUELLER, 1969]:

- (a) UT reflects the actual non-uniform rotation of the earth. It is burdened with the effect of polar motion in as much as the local astronomical meridians defining the UT (through  $\Lambda$ ) are slightly displaced (see §15.2).
- (b) UT1, again, depicts the actual non-uniform rotation of the earth, but there is no effect of polar motion.
- (c) UTC is the broadcast (transmitted) time that represents a smooth rotation of the earth; it is not, however, corrected for propagation delays in the transmitter.
- (d) UT2 is the smoothest UT with all the corrections applied.

Clearly, it is UT1 that corresponds to the GAST needed for transforming the TRA to the IT system at any given instant. The corrections needed to obtain the GAST from the broadcast time are given in any standard textbook on astronomy. Conversion from *atomic time* and *ephemeris time* (time scales not based on the earth's rotation rate) to any of the UTs is also possible, and the conversion is done through corrections published weekly by the BIH (cf. §5.4). The UTC is kept to within 0.7 s of the UT1 by the introduction of leap seconds.

It may be useful to close by recapitulating the coordinate systems introduced in this section and the transformations between them. This is done in FIG. 12.

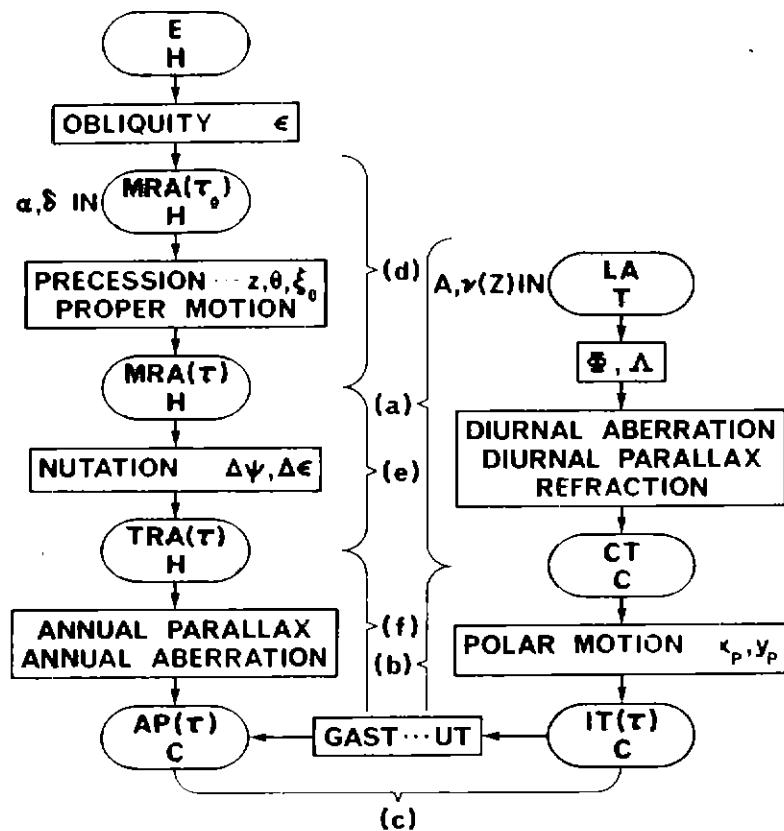


FIG. 15.12. Updating of star coordinates  $\alpha, \delta$  and reduction of observations  $A, v, (Z)$ .

## **15.2. Astronomical positioning**

As stated above, astronomical positioning is understood to be the determination of the astronomical latitude  $\Phi$  and longitude  $\Lambda$  of a point by means of particular observations  $\mathcal{I}$  to stars. The determination of the astronomical azimuth  $A$  to another point (by astronomical means) has traditionally been considered an integral part of this task, and it is thus treated here as well. The most widespread mathematical models linking the observables  $\mathcal{I}$  with the astronomical positions or azimuth fall into three classes:

- (a) latitude models:  $f_1(\Phi, l) = \mathbf{0}$ ,
  - (b) longitude models:  $f_2(\Lambda, l) = \mathbf{0}$ ,
  - (c) azimuth models:  $f_3(A, l) = \mathbf{0}$ .

Clearly,  $\Phi$  and  $\Lambda$  can be determined separately or together depending on the form of the model and the observables  $I$  involved. Observations can be collected with different accuracies, and one speaks of first, second, etc. order of accuracy, much the same way as for orders of geodetic networks (see §7.1). Implicitly involved in all the above models are the coordinates of the observed stars; throughout this section, we will consider these coordinates  $(\alpha, \delta)$  to be known in the AP system, obtained from, e.g., the APFS (see §15.1).

Most commonly used latitude models require the observations of two quantities: the zenith distance  $Z$  (or vertical angle  $\nu$ ), and the hour angle  $h$  of a star. The latter,

however, cannot be measured directly; it is determined as the sum of  $\alpha$ , obtained for the particular star from a star catalogue, and of measured time (cf. FIG. 6). Longitude models need only the knowledge of  $h$ , i.e., the time and the  $\alpha$  of the star used. Simultaneous models for  $\Phi$  and  $\Lambda$  are characterized by requiring the measurements of the zenith distances ( $Z_1$  and  $Z_2$ ) and hour angles ( $h_1$  and  $h_2$ ) to at least two stars. Azimuth models use either  $Z$  or  $h$  and the horizontal angle between the selected star and the desired point. Models also exist which do not fit into this classification; for these, the reader is advised to consult, e.g., MUELLER [1969].

The basic instrument required for measuring zenith distances is the universal theodolite. Specialized forms of theodolites are employed for very high accuracies. Other instruments may also be used; these are mentioned within the context of the mathematical models. For precise timing, one needs a chronometer (timepiece) and an HF radio receiver equipped with an amplifier and chronograph. Auxiliary equipment to measure air temperature and pressure is also used as these are needed for determining the vertical refraction correction to zenith distances.

(a) The *latitude mathematical model* is obtained from the transformation between the AP and LA systems. One gets

$$\bar{e}^{LA} = \mathbf{R}_3(\pi) \mathbf{R}_2\left(\frac{1}{2}\pi - \Phi\right) \mathbf{P}_2 \mathbf{R}_3(LAST) \bar{e}^{AP} \quad (15.13)$$

by substituting in the respective Cartesian components from (1) and (4). After some development, the equation for the third component of (4) is obtained in the following form:

$$\boxed{\sin \Phi \sin \delta + \cos \Phi \cos \delta \cos h - \cos Z = 0.} \quad (15.14)$$

Before the observed zenith distance  $Z$  is inserted in the above model, or any model for that matter, it should be corrected for the astronomical refraction effect. The *zenith distance astronomical refraction correction*  $\Delta Z$  follows from (9.12); namely,

$$\Delta Z = -\tan Z \int_1^{n_0} \frac{dn}{n}, \quad (15.15)$$

where the integration is made from outside the atmosphere (where  $n = 1$ ) to the observer (where  $n = n_0$ ). Since  $n$  as a function of time and location is not known exactly, the above integral cannot be evaluated exactly either. A practical solution can be arrived at by adopting a model for the composition of the atmosphere. Using this model, the integral can be approximated, usually in the form of a series. For details see, e.g., GARFINKEL [1944] and SAASTAMOINEN [1973]. The final result of this approximation, valid for  $Z \leq 75^\circ$ , is of the form

$$\Delta Z = C_B C_T \Delta Z^m, \quad (15.16)$$

where  $\Delta Z^m$  is the value of the mean refraction correction evaluated for some model

atmosphere, and  $C_B, C_T$  are coefficients that account for the actual barometric pressure and temperature at the time of observation. Various tables exist for these quantities: see, e.g., HOSKINSON AND DUERKSEN [1952] and MUELLER [1969].

The question that now has to be asked is: Where should the star needed for the latitude determination be located to get the most accurate result? To answer it, let us first derive the expression for the azimuth  $A$  of a star as a function of  $\Phi$ ,  $\delta$ , and  $h$ . This expression is obtained from (1), (4), and (6) as

$$\tan A = \frac{\sin h}{\sin \Phi \cos h - \tan \delta \cos \Phi}. \quad (15.17)$$

Taking the total differential of the second equation (14) and incorporating (17), the following is obtained:

$$d\Phi = -\sec A dZ - \cos \Phi \tan A dh. \quad (15.18)$$

Since the zenith distance can be measured more accurately than the time, i.e.,  $dZ < dh$ , we find that the optimum configuration, i.e., the smallest  $d\Phi$ , occurs when  $A \rightarrow 0$  or  $\pi$ . This happens when measurements are made to stars either transiting the observer's meridian (in either upper or lower *culmination*) or to circumpolar stars. Under these circumstances, an error in timing,  $dh$ , will have a negligible effect on  $\Phi$ , and  $d\Phi \rightarrow -dZ$ .

The remaining troublesome effect is that of the unaccounted for, residual astronomical refraction on the zenith distance. It can nearly be eliminated by observing pairs of stars symmetrical with respect to and close to the zenith. Residual refraction in zenith distances of such stars nearby cancel each other; the cancellation occurs because of the symmetrical composition of the atmosphere with respect to the zenith. The method based on this fact is called the *method of latitude determination by meridian zenith distances* and is in widespread use. For a star-pair transiting the meridian in upper culmination (north and south of the zenith), the model becomes [MUELLER, 1969]

$$\Phi = \frac{1}{2}(\delta_S + \delta_N) + \frac{1}{2}(Z_S - Z_N). \quad (15.19)$$

while for a star-pair in lower culmination (north and south of the zenith), the model is

$$\Phi = \frac{1}{2}(\delta_S - \delta_N) + \frac{1}{2}(Z_S - Z_N) + \frac{1}{2}\pi. \quad (15.20)$$

First-order accuracy ( $\sigma_\phi = 0.2''$ ) can be achieved by employing the above models (19) and (20) simply by using accurate instrumentation and measuring techniques. Instead of measuring the zenith distance differences by changing the position of the telescope and subtracting the two vertical circle readings, they are measured directly by means of an impersonal micrometer mounted in the telescope of the universal theodolite. This means that the star-pair must be selected so that each star is

alternately visible in the field of view of the telescope upon rotating the instrument by  $180^\circ$ . To ensure that the telescope points properly in the meridian plane, the tilt of the horizontal axis of the telescope should be measured by means of a Horrebow level clamped to the axis while the zenith distance is being measured. An intricate *observing list* of star pairs is needed to guide the observer through the programme which, for first-order accuracy, consists of about 20 star-pairs. For lower (second) order accuracy ( $\sigma_\phi = 2''$ ), usually six to eight pairs are observed [ROBBINS, 1976]. There are other methods yielding second-order accuracy the most popular of which is latitude determination by zenith distance of Polaris. Other instruments, such as the zenith telescope, the circumzenithal, and the modern (Danjon's) astrolabe, are also used in latitude determination [MUELLER, 1969].

(b) The *longitude mathematical model* is based on the use of eqn. (12), where the GAST is obtained through the UT by synchronizing a local timepiece (e.g., a quartz crystal chronometer) with a time standard by means of HF radio time signals, and the LAST is obtained from observed quantities. Specifically, from FIG. 6,

$$\text{LAST} = h + \alpha, \quad (15.21)$$

and  $h$  is obtained from the third equation (13) as

$$\cos h = \frac{\cos Z - \sin \delta \sin \Phi}{\cos \delta \cos \Phi}, \quad (15.22)$$

so that

$$\Lambda = \arccos \frac{\cos Z - \sin \delta \sin \Phi}{\cos \delta \cos \Phi} + \alpha - \text{GAST}. \quad (15.23)$$

Clearly, the latitude  $\Phi$  has to be known and  $Z$  observed.

Let us explore the optimum configuration for determining the longitude. The total differential of the longitude is

$$d\Lambda = -\sec \Phi \cos A d\Phi - \sec \Phi \operatorname{cosec} A dZ - dT - d\Delta T. \quad (15.24)$$

There now exist two possibilities: to observe a star off the meridian, in which case  $h$  is computed from (22); or to observe a star in the observer's meridian, in which case  $h = 0^h$  or  $12^h$ . In the former situation, known as the *method of longitude determination by zenith distances*, the effect of an error  $d\Phi$  in latitude can be eliminated and the effect of  $dZ$  minimized by observing in the prime vertical plane where  $A = \frac{1}{2}\pi$  or  $\frac{3}{2}\pi$ . To minimize refraction effect, east-west star-pairs of the same altitude are sought symmetrical with respect to the meridian. This configuration is very popular since there is an abundance of stars available for observations. The only drawback is again the preparation of the intricate observing lists (see THORSON [1965]). Timing error  $dT$  (reaction time) cannot be minimized by using a special observing technique, nor can a systematic error  $d\Delta T$  in timing. The observer's reaction time, called the *personal equation*, is derived from observing at a station of known longitude.

The *method of longitude determination by transit times* is the name given to the technique in which stars are observed transiting the meridian. The same model as above is used; namely, (23). The main limitation of the accuracy of this method is the inability to set the instrument exactly in the meridian. This effect is determined from special observations that are taken before the longitude determination, e.g., by tracking a low altitude star until the computed epoch of its culmination is reached. Also, a stride level is used to measure the inclination of the telescope's horizontal axis. This method is more straightforward than the one above yet yields an equivalent accuracy once proper precautions are taken. Other methods of longitude determination, as well as some combined latitude-longitude methods, may be found in MUELLER [1969] together with a description of other instruments (e.g., meridian circle) that can be used for longitude determination.

(c) The *astronomical azimuth mathematical model* uses eqn. (17). There, the latitude  $\Phi$  must be known, and the hour angle  $h$  is obtained from (12) and (21), taking  $\Lambda^{IT}$  to be known and the GAST as being derived from the UT. The horizontal angle between the star  $S$  and the desired terrestrial point is measured directly by a universal theodolite (first order) or a geodetic theodolite (second order) and added to the azimuth  $A$  of  $S$ .

The optimum observing configuration is found by taking the total differential of eqn. (17); namely,

$$dA = \sin A \cot Z d\Phi + \cos \Phi (\tan \Phi - \cos A \cot Z) dh. \quad (15.25)$$

Clearly, the effect of  $d\Phi$  is minimized by observing close to the meridian or to the horizon. The error in timing ( $dh$ ) is eliminated when

$$\tan \Phi = \cos A \cot Z, \quad (15.26)$$

which, in astronomy, is known as the condition for the *elongation* of the star, when the astronomical meridian and vertical circle of the star are perpendicular to each other. Obviously, both of the above conditions cannot be met simultaneously.

The answer to the minimization of both effects is again sought in observing a star-pair which fulfills certain conditions. Based on these conditions, STOCH [1963] has prepared charts that are designed to help with star selection. MUELLER [1969] provides guidelines for the *method of azimuth determination by hour angles of stars near culmination* as follows:

–For  $\Phi < 15^\circ$ , the two stars to be observed near the observer's meridian should have zenith distances

$$Z_1 = 75^\circ, \quad Z_2 = \text{arccot}(2 \tan \Phi + \cot 75^\circ). \quad (15.27)$$

–For  $\Phi > 15^\circ$ , Polaris should be observed at every hour angle. The alternative is to observe two north stars near the meridian, above and below the pole, whose zenith distances satisfy the following equation:

$$\cot Z_1 + \cot Z_2 = 2 \tan \Phi. \quad (15.28)$$

When the latitude is accurately known, the only error to worry about is the error in timing, and this is eliminated by observing at the elongation. Also,  $Z$  is kept as large as possible to minimize the effect of instrumental errors. This is known as the *method of azimuth determination by hour angles of stars near elongation*.

To finish this section, let us point out that all the methods shown above give astronomical quantities ( $\Phi$ ,  $\Lambda$ , and  $A$  of a terrestrial point) in the IT coordinate system, because the observations made in the LA system were transformed directly into the AP system using (6) and (9). If their counterparts,  $\Phi^{CT}$ ,  $\Lambda^{CT}$ ,  $A^{CT}$ , in the CT system are needed, the transformation inverse to (8) is used. Realizing that a unit vector in the CT system is given as

$$\bar{e}^{CT} = \begin{bmatrix} \cos \Phi^{CT} \cos \Lambda^{CT} \\ \cos \Phi^{CT} \sin \Lambda^{CT} \\ \sin \Phi^{CT} \end{bmatrix}, \quad (15.29)$$

and similarly for the IT system, this transformation yields, for instance for the third component,

$$\sin \Phi^{CT} = \sin \Phi^{IT} + \cos \Phi^{IT} (y_p \sin \Lambda^{IT} - x_p \cos \Lambda^{IT}). \quad (15.30)$$

Division by  $\cos \Phi^{IT}$  and development of  $\sin \Phi^{CT}$  into a Taylor series around  $\Phi^{IT}$  yields

$$\boxed{\Phi^{CT} \doteq \Phi^{IT} + y_p \sin \Lambda - x_p \cos \Lambda.} \quad (15.31)$$

The equation for the longitude follows similarly from the transformation of the first two components:

$$\boxed{\Lambda^{CT} \doteq \Lambda^{IT} - (x_p \sin \Lambda + y_p \cos \Lambda) \tan \Phi.} \quad (15.32)$$

The corresponding equation for the azimuth is derived from the reasoning that the observer's astronomical meridian is displaced by the polar motion. This leads to the following simple result [MUELLER, 1969]:

$$\boxed{A^{CT} = A^{IT} - (x_p \sin \Lambda + y_p \cos \Lambda) \sec \Phi.} \quad (15.33)$$

Note that the superscripts CT and IT have been dropped in some terms in the above three equations as it is immaterial which values are used to evaluate the corrective terms. In the rest of this chapter, the superscript CT will be similarly dropped.

### 15.3. Satellite positioning

Satellite methods utilize artificial earth satellites to which ranges, range differences, directions, and combinations thereof, are measured for the purpose of determining the coordinates of the observing, also known as tracking, station.

*Satellite point positioning*, in contrast with the other satellite methods to be discussed in §16.1 and §17.3, requires the coordinates (position) of the satellite to be known. These are usually given in the orbital coordinate system.

In defining this system, we imagine that in the first approximation the satellite obeys Kepler's laws (cf. §5.1) and moves along an *orbital ellipse* with the earth's centre of mass at one of the foci (cf. FIG. 13). When the satellite's orbit is close to the earth, it is perturbed by the irregularities of the earth's gravitational field and deviates from a plane ellipse—see Chapter 23. Nevertheless, it is convenient to regard the orbit to be, in the first approximation, Keplerian, i.e., planar and elliptical, and treat the perturbations as temporal variations of the six elements describing such a *Keplerian motion*. Thus, in our development, these orbital elements will be functions of time.

The point of the satellite's closest approach to the earth is called the *perigee*, and the point of farthest recession the *apogee*. Both the perigee and the apogee lie at the ends of the major axis of the orbital ellipse, called the *line of apsides*. The size and shape of the orbital ellipse are usually defined using the major semi-axis  $a_0$  and eccentricity  $e$ , much the same way as for the meridian ellipse in §7.3. The relation between these quantities and  $b$ , the minor semi-axis of the ellipse, is given by (7.11).

Now, consider the satellite to be at a point  $S$  on the orbital ellipse. The angular distance between the perigee and  $S$  is called the satellite *anomaly*. There are three kinds of anomalies in use. The *true anomaly*  $f$  (cf. FIG. 13) is the angle between the line of apsides and the line joining the centre of mass of the earth with the satellite, reckoned from the perigee in an anticlockwise direction when viewed from the NCP or from the vernal point for orbital planes that contain the NCP and SCP (polar orbits). The *eccentric anomaly*  $E$  is the angle between the line of apsides and the line joining the geometrical centre of the ellipse with the projection of the satellite  $S'$  on the concentric circle of radius  $a_0$ . The *mean anomaly*  $\mu$  (not shown in FIG. 13) is the true anomaly corresponding to the motion of an imaginary satellite of uniform

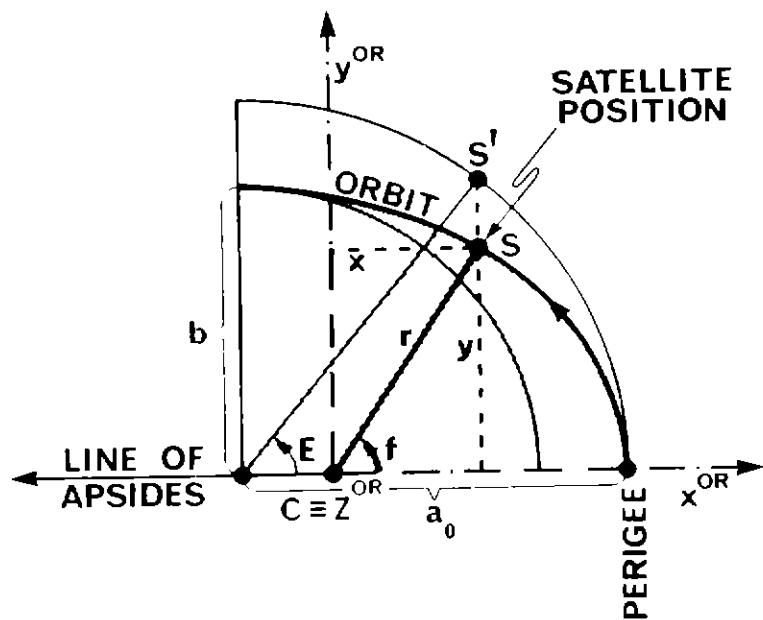


FIG. 15.13. One-quarter of a satellite orbital ellipse

angular velocity;  $\mu = 0$  at the perigee and then increases linearly with time at a rate of  $2\pi$  per revolution.

After some development, the relation between the true and eccentric anomalies is obtained from FIG. 13 as

$$\tan f = \frac{(1 - e^2)^{1/2} \sin E}{\cos E - e}. \quad (15.34)$$

The relation between the eccentric anomaly  $E$  and the mean anomaly  $\mu$  is given by *Kepler's equation* [KAULA, 1966]:

$$\mu = E - e \sin E. \quad (15.35)$$

Often the mean anomaly  $\mu$  is given, and it is necessary to find the eccentric anomaly  $E$  from (35). This can be done by iterations or by developing  $\sin E$  into a power series and inverting the equation [BROUWER AND CLEMENCE, 1961].

Now we can define the *orbital coordinate system* (OR) as follows (cf. FIG. 13): the origin is at the earth's centre of mass  $C$ , the  $x^{\text{OR}}$ -axis coincides with the line of apsides, the  $y^{\text{OR}}$ -axis corresponds to  $f = \pi/2$ , and the  $z^{\text{OR}}$ -axis completes the right-handed system. Thus the instantaneous position vector of a satellite is given by

$$\bar{r}^{\text{OR}} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}^{\text{OR}} = r \begin{bmatrix} \cos f \\ \sin f \\ 0 \end{bmatrix} = \begin{bmatrix} a_0(\cos E - e) \\ a_0(1 - e^2)^{1/2} \sin E \\ 0 \end{bmatrix}, \quad (15.36)$$

where  $r, f, a_0, e, E$  generally vary with time.

The orientation of the OR system with respect to another system has to be fixed to enable us to transform one into the other, and for this three more parameters are needed. The usual way of selecting these parameters is shown in FIG. 14. The orbital plane is extended to intersect the celestial sphere, and its trace, the projected orbit, intersects the celestial equator at the *ascending node*, the point where the satellite crosses the equator from south to north. Analogously, the *descending node* is the point of crossing from north to south (cf. §5.2). The angle between the celestial equator and the orbital half plane that contains the part of the orbit stretching from the ascending to the descending nodes is the *inclination*  $i$ . The angle between the ascending node branch of the nodal line and the line of apsides, reckoned anticlockwise looking toward the origin, is the *argument of perigee*  $\omega$ . The angle between  $x^{\text{AP}}$  (true vernal equinox) and the nodal line, measured anticlockwise from the  $+Z^{\text{OR}}$  axis in the equatorial plane, is the *right ascension of the ascending node*  $\Omega$ .

These three quantities, along with those defining the orbital ellipse and the motion of the satellite in the orbit, constitute the six *Keplerian orbital elements*. Because they

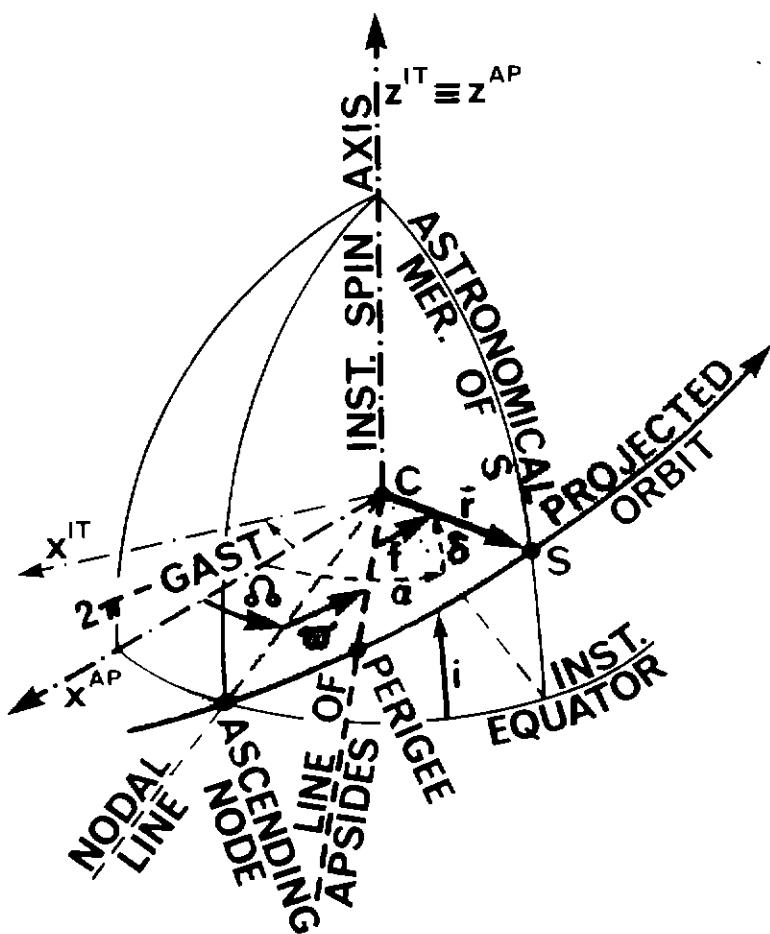


FIG. 15.14. Keplerian orbital elements.

represent a very important parameterization of the orbit, we list them below:

$a_0$ major semi-axis	size and shape of the orbit,
$e$ eccentricity	
$\omega$ argument of perigee	position of the orbit in the AP system,
$\alpha$ right ascension of ascending node	
$i$ inclination	position of the satellite in orbit.
$\mu$ mean (or other) anomaly	

There are other alternative and equivalent parameterizations of the orbit in use. For example, the geocentric Cartesian representation  $(x, y, z, \dot{x}, \dot{y}, \dot{z})$ , which describes the position of the satellite  $(x, y, z)$  along with its velocity vector  $(\dot{x}, \dot{y}, \dot{z})$  at a given epoch, will be used in §17.3. Another Cartesian system is shown in §23.2.

Point positioning requires the transformation of positions of the satellite, usually predicted ahead of time and broadcast by the satellite, from the OR system to a terrestrial system (usually the CT), where they are used to compute the coordinates

of the observing station. It is helpful to realize that the orbital plane does not rotate with the earth but remains, in the first approximation, fixed in the AP system, and that the OR and AP systems both have their origin at the earth's centre of mass (cf. FIG. 14). The transformation from the OR to CT system proceeds in three steps: OR  $\rightarrow$  AP, AP  $\rightarrow$  IT, and IT  $\rightarrow$  CT. The first and second steps should be clear from FIG. 14. The third step is given by the inverse of (8). When put together, the complete transformation reads

$$\bar{r}^{CT} = \mathbf{R}_2(-x_p) \mathbf{R}_1(-y_p) \mathbf{R}_3(GAST) \mathbf{R}_3(-\alpha) \mathbf{R}_1(-i) \mathbf{R}_3(-\omega) \bar{r}^{OR}. \quad (15.37)$$

Note that  $\bar{r}^{OR}$  is given by (36). It is thus not difficult to see that even the CT coordinates of the satellite are functions of time  $\tau$ , and we speak about the time-varying position, or *ephemeris*, of the satellite. The way the ephemeris is generated—through prediction or interpolation—will be shown in Chapter 23.

With the satellite position expressed in the CT coordinate system, we can formulate the various models for positioning:

(a) The *range mathematical model* can be written, for example, as (see FIG. 15)

$$\bar{e}_i^j \bar{r}_i = \bar{e}_i^j \bar{r}^j - \rho_i^j, \quad (15.38)$$

where  $\rho_i^j = \rho(\bar{r}_i, \bar{r}^j)$  is the measured range from tracking station  $P_j$  to satellite position  $S_j$ ;  $(x^j, y^j, z^j) = \bar{r}^j$  (obtained from (37)) are the known Cartesian coordinates of the satellite at time  $\tau_j$ ; and  $(x_i, y_i, z_i) = \bar{r}_i$  are the unknown Cartesian

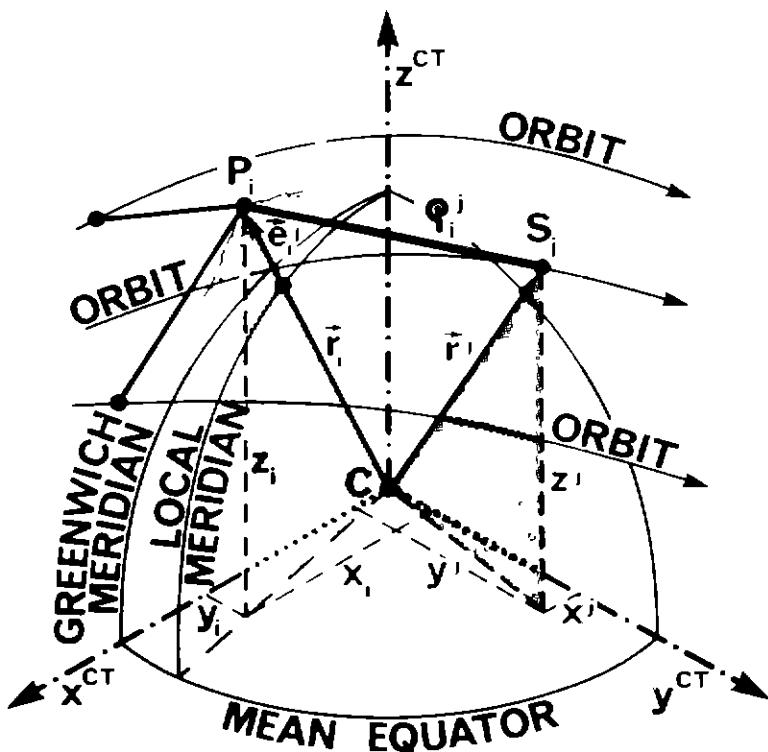


FIG. 15.15. Ranging to satellites.

coordinates for the tracking station, all in the CT system. For each range measurement, one such vector equation can be written, and three such (linearly independent) equations give a unique solution for the three unknown coordinates  $\bar{r}_i$ . Configurations of satellite positions that do not give a unique solution, called *critical configurations*, were studied extensively by BLAHA [1971b]. Observations of more than three ranges lead to an overdetermined set of equations, which can be treated by the method of least squares, as discussed in Part III.

It is of interest to observe that when the satellite ephemeris is specified in the OR (rather than the CT) system, then the coordinates  $x_p, y_p$  of the instantaneous pole or the GAST or both of the may also be treated as unknown parameters and may be determined together with  $\bar{r}_i$ .

Range measurements are obtained by timing the travel time of electromagnetic waves between the tracking station and the satellite (see §9.2). The source can be either earthbound while the satellite carries a retroreflector, or satellite borne. To date, the most accurately timable source of an electromagnetic signal devised is laser which emits short pulses of monochromatic light of a duration of a few nanoseconds (ns); it is normally earthbound. Presently the round trip travel time of a laser pulse to a reflecting satellite, such as the LAGEOS—*satellite laser ranging* (SLR)—can be determined electronically to an accuracy of 0.15 to 0.30 ns which corresponds to an accuracy of about 5 to 10 cm in the range  $\rho$  [SMITH ET AL., 1979]. The limiting factor is the inability of the electronic circuitry to more accurately detect either the ends or the centre of the returned pulse. It should be pointed out that ranging is by no means confined to artificial satellites. When several retroreflectors were placed on the lunar surface, *lunar laser ranging* became a practical and viable alternative.

A second mode of ranging, used at present, times radio signals. The radio signal used for the *satellite radio ranging* (SRR) must be of a frequency higher than 30 MHz to penetrate through the ionosphere (cf. §9.2). It also should be as coherent as possible, thus requiring the use of highly stable oscillators. Cesium (Ce) and rubidium (Rb) atomic oscillators are used for this purpose in the satellites of the NAVSTAR *Global Positioning System* (GPS). Because the emitted radiowave is continuous, the timing of the travel time must be arranged differently from the timing of discrete laser pulses. The timing is done by the ground *satellite receiver* whose clock (oscillator) must be not only highly accurate but also synchronized with the satellite oscillators so that it can associate the reception of timing marks with the time elapsed between their emission and reception. This synchronization usually cannot be done accurately enough, and the resulting *clock offset*  $\Delta t_s$ , really the (approximately) constant difference between the satellite time and the ground time, has to be included as an additional unknown in the model (38) which then acquires a term  $-c\Delta t_s$  on the left-hand side, where  $c$  is the speed of light. Under these conditions, clearly more than three ranges have to be observed to yield a solution  $(\bar{r}_i, \Delta t_s)$  of the model.

Since the GPS is quickly becoming the most important geodetic positioning system for the near future, let us describe it here in some detail. In its final constellation, scheduled for the late 1980s, the system will have 18 satellites, 6 equidistantly spaced in each of the three orbital planes with identical inclinations of

$i \doteq 55^\circ$  [JORGENSEN, 1980]. The orbital period of 12 hours and the altitude of 20 000 km will be common to all 18 satellites. This constellation has been chosen so that at least 4 satellites are visible for 24 hours a day from any point on the surface of the earth.

The GPS satellites emit highly coherent cross-polarized carrier signals on frequencies of  $L_1 = 1227.60$  MHz and  $L_2 = 1575.42$  MHz. These, in turn, are modulated by two pseudo-random sequences of zeros and ones, called P- and C/A-codes (C/A-code on  $L_1$  only), of frequencies 10.23 MHz and 1.023 MHz. The ground receiver should then be capable of generating at least one of these codes and, by matching the so-generated sequence with the incoming sequence, to measure the difference between the emission time  $\tau$  and the reception time  $t$ . This difference, corrected for the above mentioned clock offset and multiplied by the speed of propagation, gives the desired range. (Let us just mention in passing that ranges are not the only observables obtained from the GPS. Other modes can be used as will be discussed in §16.1.)

The last modulation of the carrier signals of each of the satellites is by a low frequency (50 Hz) stream of data describing the ephemeris of that particular satellite, called the *broadcast ephemeris*. The ground receiver may be designed to decipher this message, which also contains information about the clock and the general state of health of the emitting satellite, and convert it to the appropriate satellite position  $\vec{r}'$  at the time  $\tau$ , of the range measurement. Alternatively, a more *precise ephemeris* may be used instead.

Before the observed ranges  $\rho$  are introduced into the mathematical model (38), they must be corrected for the effect of astronomical refraction. The general formula for the *range refraction correction*, also called the (*propagation*) *delay correction*, takes on the form (cf. (9.9)):

$$O\rho = \int_{\mathcal{C}} (n - 1) dS, \quad (15.39)$$

where  $n$  is the index of refraction along the path  $\mathcal{C}$  between  $P_i$  and  $S_j$  (eqn. (9.5)). Since the two indices of refraction for the troposphere and the ionosphere are completely different, the integration is normally broken into two separate parts. It is then usual to speak of *tropospheric delay correction*  $O\rho^T$  and *ionospheric delay correction*  $O\rho^I$ . The above integral requires that  $n$  once more be known all along the path. Several researchers have contributed toward developing a formula for the tropospheric delay correction (see, e.g., HOPFIELD [1969] and YIONOULIS [1970]), and a good review of their work can be found in WELLS [1974]. Generally, the tropospheric correction can be written as

$$O\rho^T = \sum_{i=d,w} K_i / \sin[(\nu^2 + \theta_i^2)^{1/2}], \quad (15.40)$$

where  $\nu$  is the vertical angle to the satellite. Subscript  $d$  stands for the dry component of the air; and  $K_d$  is a function of temperature  $T$ , pressure  $P$ , and the height  $H$  of the tracking station. Subscript  $w$  stands for the wet component, and  $K_w$  is again a function of  $T$ ,  $P$ ,  $H$ , and, in addition, the vapour pressure  $e$ . The corrections to the

vertical angle  $\nu$  are  $\theta_d = 2.5^\circ$  and  $\theta_w = 1.5^\circ$ . Average values may be adopted for  $K_d$ ,  $K_w$ ; for example, MOFFETT [1971] gives the following average values for a maritime climate:  $K_d = 2.31$  m and  $K_w = 0.20$  m. For a formula different from the above, see SAASTAMOINEN [1973]; any of these formulae are generally valid only for  $\nu$  greater than  $10^\circ$  and are accurate to about 0.2 m.

The ionospheric delay correction is obtained, first by expressing  $n$  in the form of a series:

$$n = 1 + \frac{c_1}{f^2} + \frac{c_2}{f^4} + \dots, \quad (15.41)$$

where  $f$  is the frequency of the signal, and  $c$  are functions of the ray's trajectory and time, independent of  $f$ . Equation (39) then yields

$$Op^I = \int_{\mathcal{C}} (n - 1) dS = \frac{b_1}{f^2} + \frac{b_2}{f^4} + \dots, \quad (15.42)$$

where the  $b$ 's are again independent of  $f$ . Note that this correction, being a function of frequency, can be evaluated if ranges are measured simultaneously on two different frequencies, as is the case with the NAVSTAR system. We shall return to this correction later.

The error in position determination using range data can be broken down into two main components: the error in range and the error in ephemeris, also called the orbital error, or error in the satellite position. As an example, for the GPS, the error in the broadcast ephemeris is about  $\sigma_r = 2$  m [ANDERLE, 1980], and the errors in ranges depend on which 'code' is used for ranging. P-code ranges should have standard deviations of about 1.5 m, while C/A-code ranges are good to better than 10 metres. A position determined by P-code ranging for a few hours should be accurate to about 0.5 m (one standard deviation, internal consistency) [WELLS ET AL., 1981]. On the other hand, the standard deviation of SLR-determined positions from observations spread over many months is reported to be about 2 to 3 cm [CHRISTODOULIDIS AND SMITH, 1983]. Comparable data for the GPS could not be located.

(b) Logically, the next model that should be treated here is the *direction mathematical model*. To obtain directions to a satellite, the satellite is photographed against a background of known stars. Then star and satellite positions are identified on the photographic plate, and measurements are made of the location of the satellite image relative to the images of known stars. With the  $\alpha$  and  $\delta$  of the stars known, the  $\alpha$  and  $\delta$  of the satellites can be estimated. This seemingly simple procedure has several problems that limit its accuracy and make it, in fact, quite complicated. The most troublesome is the uncertainty in modelling the camera lens distortion which can only be done by introducing many redundant observations. The accuracy of the catalogued star positions presents additional limitations. The effect of atmospheric refraction (eqn. (15)) is also a problem [MUELLER, 1964]. As in the case of the range model, the orbital errors contribute further uncertainties.

It is not possible to determine the three-dimensional position of a single tracking station from direction observations alone. At least one distance either on the ground or to a satellite has to be measured to supply the missing scale. Directions will thus be considered only as part of the next model that uses simultaneously measured directions and ranges and is conceptually much simpler. Further uses of satellite directions will be dealt with in §16.1 and §17.3.

(c) The *simultaneous direction and range mathematical model* can be written as (see FIG. 16)

$$\bar{r}_i = \bar{r}^j - \bar{\rho}_i^j, \quad (15.43)$$

where  $\bar{r}_i$  and  $\bar{r}^j$  are again radius vectors of the observing point and satellite respectively. The *topocentric vector*  $\bar{\rho}_i^j$  in the CT system is given by

$$\bar{\rho}_i^{j\text{CT}} = \mathbf{R}_2(-x_p)\mathbf{R}_1(-y_p)\mathbf{R}_3(\text{GAST})\rho_i^j \begin{bmatrix} \cos \delta_j \cos \alpha_j \\ \cos \delta_j \sin \alpha_j \\ \sin \delta_j \end{bmatrix}. \quad (15.44)$$

Clearly, even with only one observation  $\bar{\rho}_i^j$ , it is possible to determine uniquely the position  $\bar{r}_i$  of the tracking station. When more than one satellite position is observed, then (43) becomes one of several observation equations, and averaging has to be used to give the three components of  $\bar{r}_i$ . In this instance, note that the model is of the explicit-in-parameters ( $x_i, y_i, z_i$ ) variety the handling of which has been discussed in Chapter 11. As this method of point positioning is not widely used, there is no good estimate available of the achievable point position accuracy.

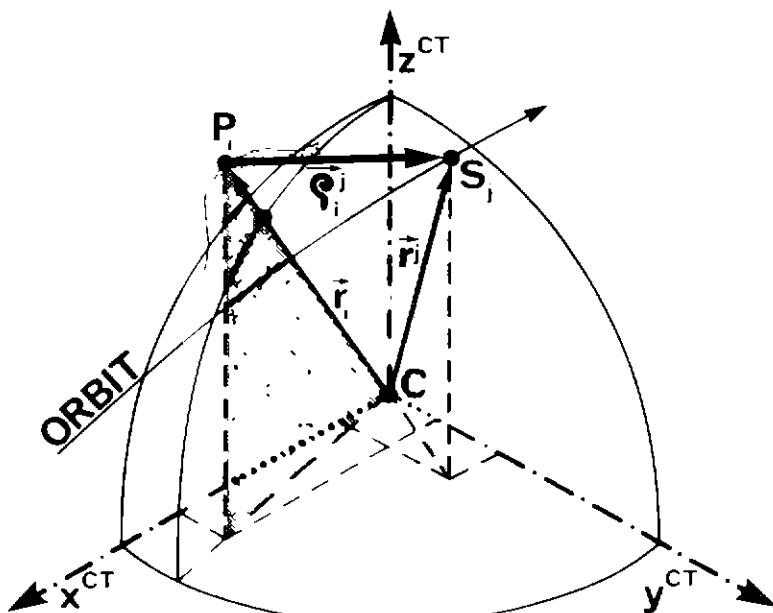


FIG. 15.16. Simultaneous ranging and directions to a satellite.

(d) When a moving object, such as an orbiting satellite, transmits an electromagnetic signal of a certain (constant) frequency  $f_T = 2\pi/\lambda_T$ , the observer receives a signal of frequency  $f_R = 2\pi/\lambda_R$  which varies with the velocity of the transmitter with respect to the receiver (see FIG. 17). Mathematically, the received wavelength  $\lambda_R$  is related to the constant transmitted wavelength  $\lambda_T$  through the *Doppler equation* [MENZEL, 1955]

$$\lambda_R = \lambda_T \left( 1 + \frac{\dot{r}}{c} \right) \left( 1 + \frac{\dot{r}^2}{c^2} \right)^{1/2}, \quad (15.45)$$

where  $c$  is the speed of light and  $\dot{r}$  is the rate of change of the range with time (see FIG. 18). Equation (45) is said to describe the *Doppler effect*.

From eqn. (45) we see that  $\lambda_R = \lambda_T$  when and only when  $\dot{r} = 0$ . This situation occurs only when the satellite moves with a velocity  $\vec{v}$  normal to the range vector  $\vec{r}$  (cf. FIG. 18). The point of the satellite orbit where the satellite velocity is normal to the satellite range vector is the *point of closest approach* (PCA). Only the frequency ( $f_T$ ) transmitted at the PCA is later received by the receiver undistorted.

The TRANSIT satellite system was designed to exploit the Doppler effect in a novel way. The system employs five satellites with nearly circular ( $e \approx 0$ ) polar ( $i = \pi/2$ ) orbits of a mean altitude of about 1074 km (i.e.,  $a_0 \approx 6371 \text{ km} + 1074 \text{ km} = 7445 \text{ km}$ ). This altitude implies the orbital period of about 107 minutes (see eqn. (5.1)) and the translational velocity  $v$  of about  $7.3 \text{ km s}^{-1}$ . With  $\dot{r}$  being always smaller than  $v$ , the relativistic effect  $(\dot{r}/c)^2$  in eqn. (45) accounts for less than  $3 \times 10^{-10}$  of the wavelength and thus is normally neglected.

Now, substituting frequencies for wavelengths in eqn. (45), we obtain

$$f_R \doteq \frac{2\pi}{\lambda_T (1 + \dot{r}/c)} \doteq f_T (1 - \dot{r}/c). \quad (15.46)$$

Then it is easy to express the range rate  $\dot{r}$  as a function of the two frequencies:

$$\dot{r} \doteq c (1 - f_R/f_T). \quad (15.47)$$

Finally, the *range difference*  $\nabla r$  between two satellite positions  $S(\tau_j) = S'$  and  $S(\tau_k) = S^k$  is given simply as

$$\nabla r \doteq \frac{c}{f_T} \int_{\tau_j}^{\tau_k} (f_T - f_R) d\tau. \quad (15.48)$$

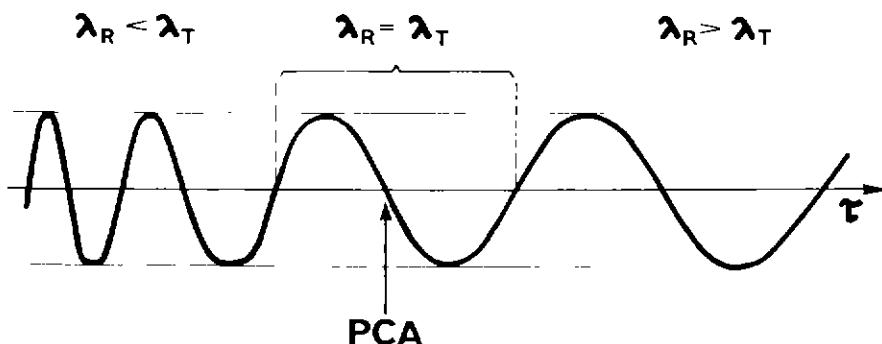
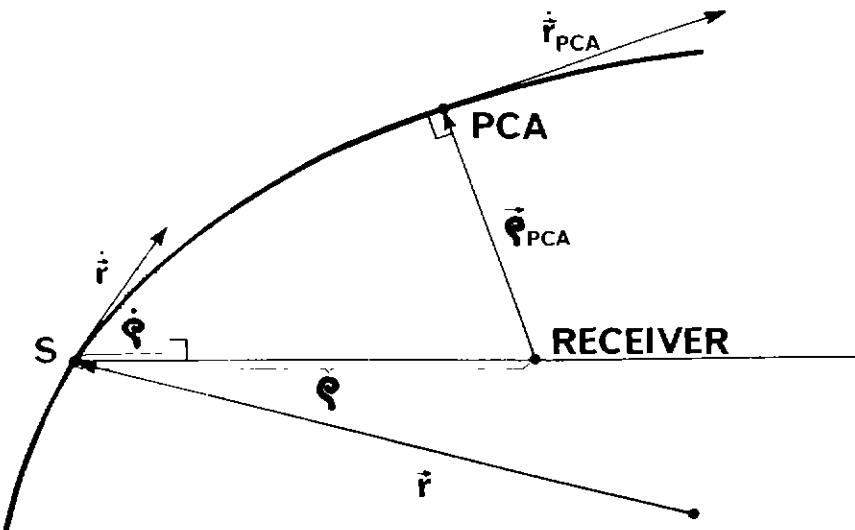


FIG. 15.17. Variation of received signal wavelength.

FIG. 15.18. The rate of change of the range  $\rho$ .

Assuming that the beat frequency  $f_T - f_R$  can be observed and integrated, how can the range differences  $\nabla\rho$  be used to obtain the position of the receiver? The observed  $\nabla\rho$ 's can be used in a scheme known as *hyperbolic positioning*, the concept of which is as follows. Suppose we know two positions  $S^1, S^2$  of the satellite and the  $\nabla\rho_i^{12}$  observed at the point  $P_i$ . From elementary geometry we know that  $P_i$  must lie on one of the hyperbolic surfaces  $\mathcal{H}_1, \mathcal{H}_2$  shown in FIG. 19, because each hyperbolic surface is a locus of a constant difference in ranges reckoned from the two foci  $S^1$  and  $S^2$ . If we then get another range difference, e.g.,  $\nabla\rho_i^{23}$ , relating  $P_i$  to another pair of hyperbolic surfaces, then  $P_i$  must lie on one of the curves made by intersecting the appropriate hyperbolic surfaces. A third range difference should then provide us with only one point at the intersection of two such curves.

The TRANSIT system is designed so that a whole string of satellite positions on one *orbital arc*, also called a *satellite pass*, is used. One satellite pass is usually taken as the visible part of the orbit that spans two successive passages through the almucantar of  $Z = 82^\circ$ . The satellite positions on each pass are spaced equidistantly in time and the used constant time interval  $\Delta\tau$  depends on the design of the receiver, the optimum being about 30 s.

To understand how the TRANSIT range differences are actually obtained, we have to know at least the basic principles on which the system works. Each satellite transmits signals on two stable crystal oscillator generated frequencies of 150 MHz and 400 MHz. These two signals are received on the receiver's omnidirectional antenna and compared to similar, internally generated frequencies. As an example, let us have a closer look at this comparison for the 400 MHz frequency. First, the transmitted frequency  $f_T$  is deliberately offset by about 32 kHz to be slightly lower (399.968 MHz) than the ground frequency  $f_G = 400$  MHz generated by the receiver's oscillator. Fig. 20 shows both the received frequency  $f_R$  (that depends on the satellite's velocity) and the situation that occurs during one satellite pass. The approximately 32 kHz offset of the transmitted frequency ensures that the difference between  $f_G$  and  $f_R$  is always positive (between 24 and 40 kHz).

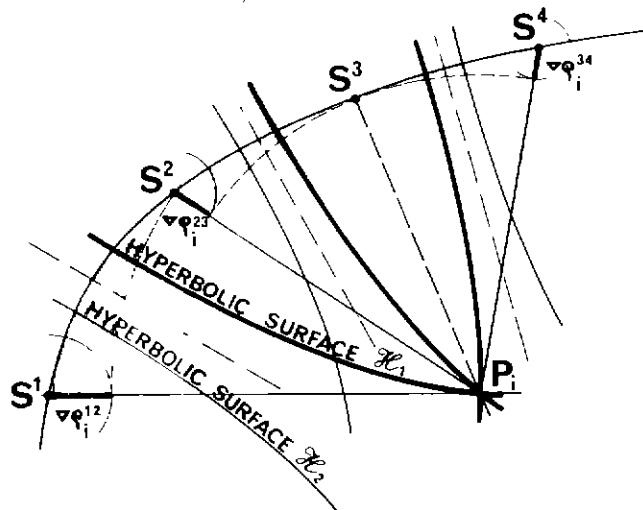


FIG. 15.19. Hyperbolic positioning.

Now, if both  $f_T$  and  $f_G$  can be regarded as sufficiently stable during one pass, which lasts somewhere between 10 and 18 minutes, then we can write

$$f_T = f_G - \Delta f, \quad (15.49)$$

where the *frequency offset*  $\Delta f$  is constant during one pass but generally changes from pass to pass around the value of 32 kHz. Substituting for  $f_T$  in eqn. (48) we obtain:

$$\nabla \rho \doteq \frac{c}{f_T} \int_{\tau_j}^{\tau_k} (f_G - f_R) d\tau - \frac{c}{f_T} \Delta f \Delta \tau, \quad (15.50)$$

where the beat (Doppler) frequency  $f_G - f_R \in \langle 24 \text{ kHz}, 40 \text{ kHz} \rangle$  is what the receiver tracks. The Doppler frequency tracked in time gives the number of cycles of the Doppler signal; this number is a unitless quantity, and it is called the *Doppler count*. Its integral, accumulated between the two time marks  $\tau_j$ ,  $\tau_k$  transmitted by the satellite, is the observable quantity recorded by the receiver and called the *integrated Doppler count*  $D$ . It is easy to see that during a 30-second interval the receiver

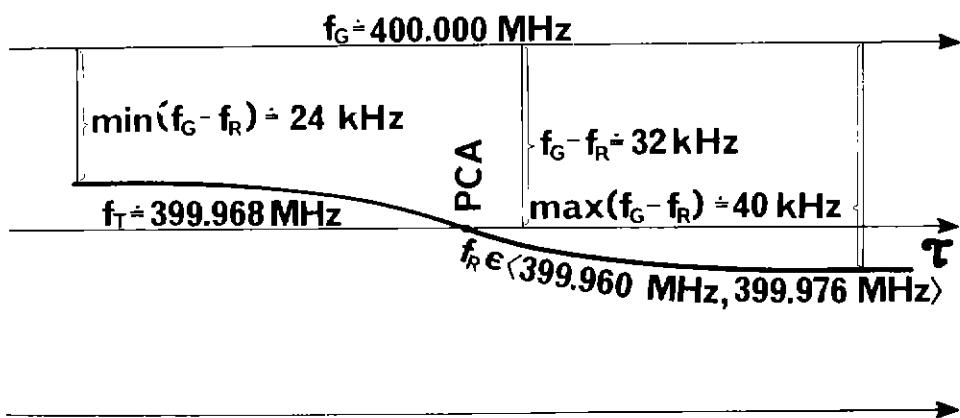


FIG. 15.20. Frequencies involved in TRANSIT positioning.

accumulates about  $10^6$  Doppler counts which are then multiplied by  $c/f_T = 75$  cm to give the first part of the range difference  $\nabla\rho$  in eqn. (50). The second part, constant for the whole pass, can be obtained only after the unknown offset  $\Delta f$  for that particular pass has been determined. An identical procedure is used for the frequency of 150 MHz.

Let us now suppose that we know the frequency offset for one pass and thus the range differences to all the adjacent ( $\Delta\tau$  apart) satellite positions on the orbital arc. Since this arc is approximately planar, all the hyperbolic surfaces (whose point intersection we seek) intersect in a family of spatial curves which will very nearly have a common tangent perpendicular to the orbital plane. Thus the determination of the position  $\bar{r}_i$  will be very weak; if the satellite pass were exactly planar, then the determination of  $\bar{r}_i$  from one pass would be impossible. At least two passes with sufficiently inclined orbital planes are therefore needed before a reasonable solution  $\bar{r}_i$  is attempted. Usually data from many passes are collected for each position determination.

The *range difference mathematical model* needed for the determination of the receiver antenna's position is obtained simply by formulating two range mathematical models (eqn. (38)) for  $P_i$  and two different satellite positions  $S^j$ ,  $S^k$ . Subtracting the first from the second, we get

$$(\bar{e}_i^k - \bar{e}_i^j) \bar{r}_i = (\bar{e}_i^k \bar{r}^k - \bar{e}_i^j \bar{r}^j) - \nabla\rho_i^{jk}. \quad (15.51)$$

Substituting for the range difference, we obtain finally

$$(\bar{e}_i^k - \bar{e}_i^j) \bar{r}_i - \frac{c}{f_T} \Delta\tau \Delta f_i^l = (\bar{e}_i^k \bar{r}^k - \bar{e}_i^j \bar{r}^j) - \frac{c}{f_T} D_i^{jk}, \quad (15.52)$$

where  $\Delta f_i^l$  refers to the  $l$ th pass observed at  $P_i$ .

In order to evaluate the position  $\bar{r}_i$  and the  $m$ -tuple of frequency offsets for as many passes, we have not only to observe the integrated Doppler counts  $D_i^{jk}$  but also have to know the positions  $\bar{r}^j$ ,  $\bar{r}^k$  of the satellites at the instants of emission of the integration time marks. These positions are evaluated either from the predicted orbital information broadcast by the satellite or from the post-fitted orbital information. The broadcast ephemeris is given in terms of coordinates very close to Keplerian orbital elements and their rates of change. Satellite positions computed from the broadcast information are good to about 25 m along the track, 15 m radially, and 5 m across the track. The other alternative is the precise ephemeris, which is about twice as accurate as the broadcast ephemeris.

It should be mentioned here that the same observable, i.e., the integrated Doppler count, can be obtained from the GPS. With the GPS satellites moving at a considerably lower angular velocity, however, there is no geometrical advantage in using the GPS in this mode [VANÍČEK ET AL., 1984].

Before the Doppler counts  $D_i^{jk}$  are inserted in the model, the tropospheric and ionospheric delay corrections must again be applied. Utilizing (52), the *tropospheric correction to Doppler count*  $D_i^{jk}$  is found to be

$$OD_i^{jkT} = \frac{f_G}{c} (\delta\rho_i^{kT} - \delta\rho_i^{jT}), \quad (15.53)$$

where  $\delta\rho_i^{kT}$  and  $\delta\rho_i^{jT}$  are the respective tropospheric corrections to the two ranges  $\rho_i^k$  and  $\rho_i^j$  (see (40)). The *ionospheric correction to Doppler count*  $D_i^{jk}$  follows from (42). For each of the two Doppler counts observed on the two frequencies  $f_1$  and  $f_2$ , we can write

$$\left. \begin{aligned} D(f_1) &= D(\text{vac}) + \frac{a_1}{f_1}, \\ D(f_2) &= KD(\text{vac}) + \frac{a_1}{f_2} = KD(\text{vac}) + \frac{1}{K} \frac{a_1}{f_1}, \end{aligned} \right\} \quad (15.54)$$

where  $D(\text{vac})$  is the correct count,  $K = f_2/f_1$ , and  $a_1$  is a constant. For the TRANSIT system,  $K = \frac{3}{8}$  (see, e.g., WELLS [1974]) and

$$OD^1 = D(400) - D(\text{vac}) = \frac{3}{8}(D(150) - \frac{3}{8}D(400)), \quad (15.55)$$

where  $D(150)$  and  $D(400)$  are the observed Doppler counts on the two frequency channels.

An important point to note is that continuous Doppler measurements are correlated since they are obtained from the differences between accumulated counts at two subsequent instants  $\tau_j$  and  $\tau_k$ . In other words, since any two adjacent Doppler counts have one accumulated count in common, they are correlated through this value. This means that the covariance matrix of observations will be tridiagonal—see §3.1—[KRAKIWSKY ET AL., 1972]. It has been shown that this correlation can be removed simply by transforming the range difference equation to a range equation [BROWN, 1970]. In this mode, the tracking station to satellite range  $\rho_i^0$  encountered at the beginning  $\tau_0$  of counting becomes an additional unknown. The accumulated Doppler counts  $D_i^{01}, D_i^{02}, \dots$  at the subsequent epochs  $\tau_1, \tau_2, \dots$  (see FIG. 21) yield uncorrelated ranges for these epochs (cf. (50))

$$\rho_i^k = \rho_i^0 + \frac{c}{f_T} [D_i^{0k} - \Delta f_i \Delta \tau], \quad (15.56)$$

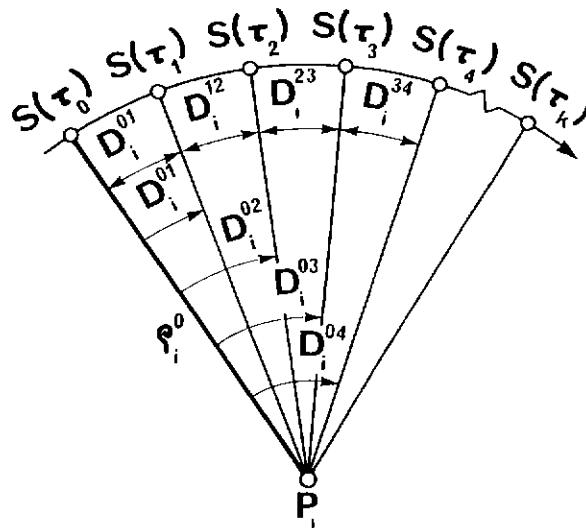


FIG. 15.21. Correlation between Doppler counts.

and has the advantage that the covariance matrix of observations, i.e., accumulated Doppler counts  $D_i^{0k}$ , is diagonal.

Clearly, the achievable accuracy in  $\bar{r}_i$  depends on the accuracy of the orbit ( $\bar{r}^k, \bar{r}'$ ) as well as the accuracy of the observed Doppler count  $D_i^{jk}$ . Considering again the TRANSIT system, the orbital accuracy is of the order of several metres [ANDERLE, 1974]. The original Doppler satellite receivers (with  $\Delta\tau = 120$  s) record  $D_i^{jk}$  to one cycle, which corresponds to an accuracy of 0.75 m in  $\nabla\rho_i^{jk}$  for  $f_s = 400$  MHz. The newest generation of receivers possesses a more refined counter, and  $D_i^{jk}$  can now be determined to 0.4 cycle and less and could give an accuracy in  $\nabla\rho_i^{jk}$  of about 0.3 m [KOUBA, 1980] or better, if it were not for the refraction. All in all, the point positions from this system when about 50 passes are used are accurate to about 1 m for the broadcast ephemerides and about 20 cm for the precise ephemerides [KRAKIWSKY ET AL., 1972]. Point positioning accuracy of the range difference technique has not improved much since 1972.

#### 15.4. Transformations of terrestrial positions

It is often necessary to transform the position of a point, normally located on the earth's surface, from one coordinate system into another. It is natural to formulate a mathematical model and solve for the coordinates in one particular coordinate system. Later, one may want to refer the coordinates of the same point to another coordinate system, and that is where the necessity for a point transformation arises. In §15.2, we encountered one such situation in which the astronomically determined position  $\Phi, \Lambda$  was solved for in the IT system and later transformed into the CT system. Central to the point transformations treated here is the concept of the reference (biaxial) ellipsoid, with its geodetic curvilinear coordinates  $\phi, \lambda, h$  (cf. §7.1). The following topics are included in this section:

- (a) The transformation of the geodetic curvilinear coordinates  $(\phi, \lambda, h)^G$  into their representative Cartesian coordinates  $(x, y, z)^G$ —cf. §3.3—and vice versa.
- (b) The transformation of the CT coordinates  $(x, y, z)^{CT}$  into non-geocentric geodetic coordinates  $(\phi, \lambda, h)^G$  and vice versa.
- (c) The transformation of the astronomical coordinates  $(\Phi, \Lambda)$  into geodetic curvilinear coordinates  $(\phi, \lambda)$ , along with the transformation of the astronomical azimuth ( $A$ ) into geodetic azimuth ( $\alpha$ ), and the orthometric height ( $H$ ) into geodetic height ( $h$ ) and vice versa. The ways of positioning of a geodetic reference ellipsoid within the earth are discussed within these transformations.
- (d) The transformation of one triplet of geodetic curvilinear coordinates  $(\phi, \lambda, h)_1$  referred to an ellipsoid  $(a_1, b_1)$  into another triplet  $(\phi, \lambda, h)_2$  referred to another ellipsoid  $(a_2, b_2)$ .
- (e) The transformation of horizontal, geodetic curvilinear coordinates  $(\phi, \lambda)$  into map coordinates  $(x, y)^M$  and vice versa.

In the above list, there are two classes of transformation. The first class, consisting of transformation (a), is the transformation within one family of coordinate systems (see §3.3). The second is the class of transformation between families of coordinate systems with different locations and orientations.

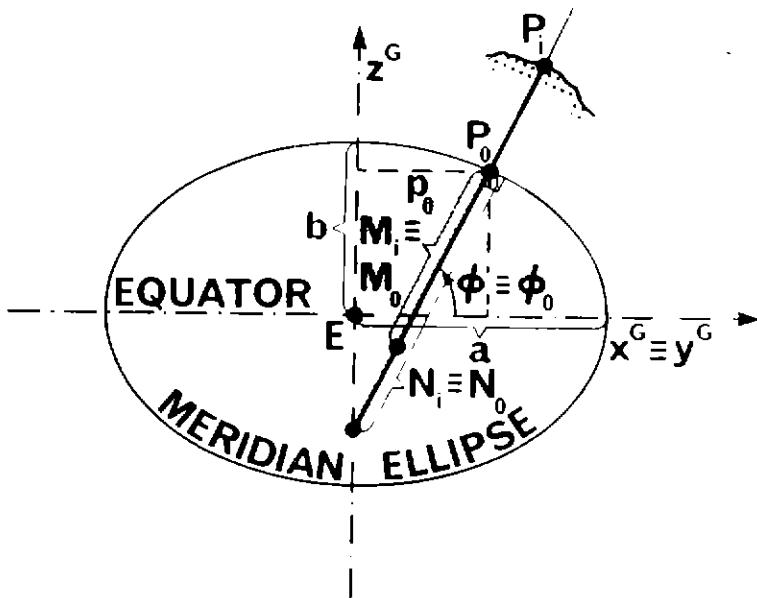


FIG. 15.22. Geometry of a biaxial ellipsoid.

Before getting into the first transformation, let us derive the expression for one more entity related to the biaxial ellipsoid. From (3.90) we obtain, for any point  $P_0$  on the ellipsoid (see FIG. 22),

$$N_0 \cos \phi_0 = p_0, \quad (15.57)$$

where  $N_0$  is called the *prime vertical radius of curvature* at  $P_0$  and can be derived from (7.10) and (7.12) as

$$N_0 = \frac{a^2}{(a^2 \cos^2 \phi_0 + b^2 \sin^2 \phi_0)^{1/2}}. \quad (15.58)$$

It is clearly akin to  $M$  given by (7.14).

(a) In the first transformation, realizing that in the representative Cartesian geodetic system (G) of coordinates (corresponding to the curvilinear system) one has

$$x_0^G = p_0 \cos \lambda_0, \quad y_0^G = p_0 \sin \lambda_0, \quad (15.59)$$

we can write the position vector of the normal projection  $P_0$  of  $P_i$  on the ellipsoid simply as

$$\begin{aligned} \bar{r}_0^G &= \bar{r}^G(\phi_0, \lambda_0) = \bar{r}^G(\phi_i, \lambda_i) \\ &= N_0 \begin{bmatrix} \cos \phi_0 \cos \lambda_0 \\ \cos \phi_0 \sin \lambda_0 \\ (b^2/a^2) \sin \phi_0 \end{bmatrix} = N_i \begin{bmatrix} \cos \phi_i \cos \lambda_i \\ \cos \phi_i \sin \lambda_i \\ (b^2/a^2) \sin \phi_i \end{bmatrix}, \end{aligned} \quad (15.60)$$

where the  $z^G$ -component is given by (7.10). To obtain the position vector of a point  $P_i$ , located above point  $P_0$  on the ellipsoid, the two constituent vectors are added as

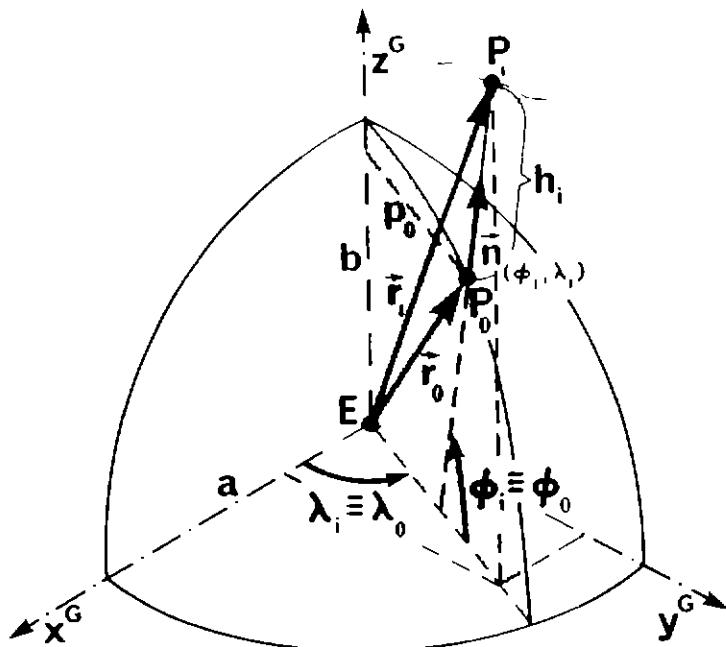


FIG. 15.23. Points on and above the reference ellipsoid.

follows (see FIG. 23):

$$\bar{r}_i^G = \bar{r}^G(\phi_i, \lambda_i) + h_i \bar{n}^G(\phi_i, \lambda_i), \quad (15.61)$$

where

$$\bar{r}^G = \bar{n}^G(\phi_i, \lambda_i) = \begin{bmatrix} \cos \phi_i \cos \lambda_i \\ \cos \phi_i \sin \lambda_i \\ \sin \phi_i \end{bmatrix} \quad (15.62)$$

is the unit vector normal to the ellipsoid at  $P_0$ , while  $h_i$  is the geodetic height of  $P_i$  (cf. §7.1). The resultant position vector then is equal to

$$\bar{r}_i^G = \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} = \begin{bmatrix} (N_i + h_i) \cos \phi_i \cos \lambda_i \\ (N_i + h_i) \cos \phi_i \sin \lambda_i \\ (N_i b^2/a^2 + h_i) \sin \phi_i \end{bmatrix}. \quad (15.63)$$

This is the transformation equation from the  $(\phi, \lambda, h)$  to the G system.

The inverse transformation can be solved by iterations or in closed form. Both approaches employ the distance  $p$  from the minor axis which, for any point  $P$  (after dropping the subscript  $i$ ), equals

$$p = (x^2 + y^2)^{1/2}, \quad (15.64)$$

or, from (63),

$$p = (N + h) \cos \phi. \quad (15.65)$$

From (63) and (7.10) we have

$$z = (N + h - e^2 N) \sin \phi, \quad (15.66)$$

and finally

$$\frac{z}{p} = \tan \phi \left( 1 - \frac{e^2 N}{N + h} \right). \quad (15.67)$$

This equation is the point of departure for both approaches.

The iterations are usually initiated by solving first for  $\phi$  from the above equation [HEISKANEN AND MORITZ, 1967]. Putting  $h=0$ , we get

$$\phi^{(0)} = \arctan \left[ \frac{z}{p} (1 - e^2)^{-1} \right]. \quad (15.68)$$

The  $k$ th iteration then consists of evaluating successively  $N^{(k)} = N(\phi^{(k-1)})$  from (58);  $h^{(k)} = h(\phi^{(k-1)}, N^{(k)})$  from (65); and  $\phi^{(k)} = \phi(N^{(k)}, h^{(k)})$  from (68). The iterations are repeated until the following inequalities are satisfied:

$$|h^{(k)} - h^{(k-1)}| < a\epsilon \quad \text{and} \quad |\phi^{(k)} - \phi^{(k-1)}| < \epsilon, \quad (15.69)$$

for some a priorily chosen value of  $\epsilon$ . Once  $\phi$  and  $h$  are found,  $\lambda$  is evaluated from either of the first two equations (63) or

$$\lambda = 2 \arctan \frac{y}{x + \sqrt{x^2 + y^2}}. \quad (15.70)$$

The closed form solution uses (65) and (66) to obtain

$$p \tan \phi - z = e^2 N \sin \phi. \quad (15.71)$$

In this equation, the only unknown is  $\phi$ ,  $N$  being a function of  $\phi$  as well. Substituting from (58), (71) changes to

$$p \tan \phi - z = \frac{ae^2 \sin \phi}{(\cos^2 \phi + (b^2/a^2) \sin^2 \phi)^{1/2}}. \quad (15.72)$$

Dividing the numerator and denominator of the right-hand side by  $\cos \phi$  and squaring the whole equation yields

$$p^2 \tan^4 \phi - 2pz \tan^3 \phi + \left( z^2 + \frac{p^2 - a^2 e^4}{1 - e^2} \right) \tan^2 \phi - \frac{2pz}{1 - e^2} \tan \phi + \frac{z^2}{1 - e^2} = 0. \quad (15.73)$$

This is a quartic (biquadratic) equation in  $\tan \phi$  in which the values of all the coefficients are known. Standard procedures for solving quartic equations exist (e.g. KORN AND KORN [1968]). Once a solution for  $\tan \phi$  is obtained,  $N$  and  $h$  are

computed from (58) and (65) respectively. Longitude  $\lambda$  follows directly from (63) or (70) thus completing the inverse transformation. PAUL [1973] has shown that the closed form approach is about 25% faster than the iterative. It should be noted that since the  $(\phi, \lambda, h)$  system is a two-parametric system of coordinates (cf. §3.3), the two parameters  $a, b$  (or  $a, e$ , or some other combination) play a role in all the above transformations.

(b) The second transformation from the CT into the G system requires a knowledge of the position and orientation of the reference ellipsoid within the earth. This task of positioning and orienting the reference ellipsoid is known as the *establishment of a horizontal geodetic datum* [YEREMEYEV AND YURKINA, 1969; MATHER, 1970; PICK ET AL., 1973]: the reference ellipsoid, defined by the values of its parameters  $a, b$ , then becomes the datum—a specific coordinate surface (cf. §3.3). The positioning of the ellipsoid requires six more parameters to eliminate its six degrees of freedom, i.e., the six ways in which the ellipsoid can move relative to the earth. Those six bring the total number of *datum parameters* to eight. Since we are interested in the transformation of the  $(\phi, \lambda, h)$  into the CT system, it is natural to specify the six datum position parameters at the earth's centre of mass—the *geocentric set of datum position parameters*—as the three CT coordinates of the ellipsoid's centre, called *datum translation components*  $x_E, y_E, z_E$ , and the three *datum misalignment angles*,  $\epsilon_x, \epsilon_y, \epsilon_z$ , required to define the misalignment between the two sets of axes (see FIG. 24). The ways datums are positioned in practice will be shown in §17.1 and §18.1. The derivation of the inverse transformation is left to the reader.

The transformation from  $(\phi, \lambda, h)^G$  into  $(x, y, z)^{CT}$  is done in two steps: first  $(\phi, \lambda, h)^G \rightarrow (x, y, z)^G$  using (63), and then  $(x, y, z)^G \rightarrow (x, y, z)^{CT}$  employing the

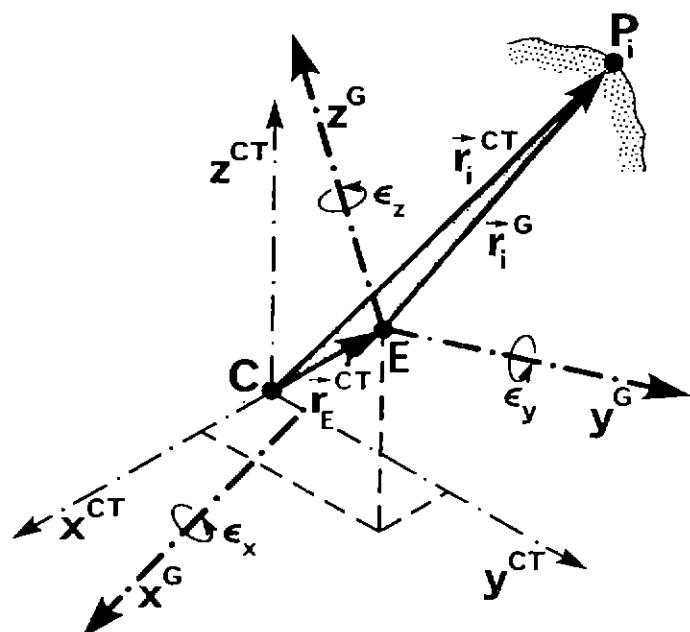


FIG. 15.24. Geocentric set of datum position parameters.

following formula:

$$\bar{r}^{\text{CT}} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}^{\text{CT}} = \mathbf{R}_1(\epsilon_x) \mathbf{R}_2(\epsilon_y) \mathbf{R}_3(\epsilon_z) \begin{bmatrix} x \\ y \\ z \end{bmatrix}^G + \begin{bmatrix} x_E \\ y_E \\ z_E \end{bmatrix}^{\text{CT}}. \quad (15.74)$$

It is obviously advantageous to have the two sets of axes parallel, i.e.,  $\epsilon_x = \epsilon_y = \epsilon_z = 0$ , so that the above equation simplifies to

$$\bar{r}^{\text{CT}} = \bar{r}^G + \bar{r}_E^{\text{CT}}. \quad (15.75)$$

In converting the CT coordinates (obtained, e.g., from satellite positioning) to the G system, the accuracy of the results depends on the accuracy of  $\bar{r}_E^{\text{CT}}$ , which is about 2 m (cf. §15.3). The present accuracy of the values of the six transformation parameters (known at least for the major datums in the world—see §7.1) contributes another 1 or 2 m to the uncertainty [THOMSON, 1976].

(c) The next transformation should enable us to transform the astronomically determined positions  $(\Phi, \Lambda)$  to geodetic positions  $(\phi, \lambda)$ . It is not as clear geometrically as the transformations treated above and requires some preliminary cognizance. In particular, the position of the G system with respect to the earth's gravity field, the framework for the  $\Phi, \Lambda$  coordinates, has to be properly understood. To explain the concepts involved, one has to use two more coordinate systems: the LA system, introduced in §15.1, and the local geodetic system.

The *local geodetic system* (LG) is defined as follows (see FIG. 25): it is topocentric (T); the  $z^{\text{LG}}$ -axis is the outward ellipsoid normal passing through T; the  $x^{\text{LG}}$ -axis is directed toward *geodetic north*, i.e., it lies in the *geodetic meridian* plane defined by the ellipsoid normal at T and the minor axis ( $z^G$ ) of the reference ellipsoid; and the  $y^{\text{LG}}$ -axis is chosen so that the system is left-handed. Note that the LG system makes angles of  $\phi$  and  $\lambda$  with the G system; clearly the relation of the LG system to the G system is analogous to the relation of the LA system to the CT system, as the reader can see by comparing FIGS. 25 and 4. The analogy between the LG and LA systems goes further: analogous to the astronomical quantities (see §15.1), here we define the *geodetic vertical angle*  $\nu'$ , the *geodetic zenith distance*  $Z'$ , and the *geodetic azimuth*  $\alpha$ , all as shown in FIG. 25. We note, however, that while the LA system is a natural system, dictated by the physical properties of the earth, the LG system is not.

Let us now examine the relations among the four coordinate systems (CT, LA, G, LG) shown in FIG. 26. A unit vector  $\bar{e}^{\text{LA}}$  can be rotated into the CT system as follows. From eqn. (74) we have

$$\bar{e}^{\text{CT}} = \mathbf{R}(\epsilon_x, \epsilon_y, \epsilon_z) \bar{e}^G. \quad (15.76)$$

Recalling the analogy between the pairs CT, LA and G, LG of systems, an equation analogous to (6) is used to transform  $\bar{e}^{\text{LG}}$  to  $\bar{e}^G$ . Finally, from FIG. 26, we get

$$\bar{e}^{\text{LG}} = \mathbf{R}_3(\Delta\alpha) \mathbf{R}_2(-\xi) \mathbf{R}_1(\eta) \bar{e}^{\text{LA}} = \mathbf{R}^T(\Delta\alpha, -\xi, \eta) \bar{e}^{\text{LA}}, \quad (15.77)$$

so that

$$\bar{e}^{\text{CT}} = \mathbf{R}(\epsilon_x, \epsilon_y, \epsilon_z) \mathbf{R}_3(\pi - \lambda) \mathbf{R}_2\left(\frac{\pi}{2} - \phi\right) \mathbf{P}_2 \mathbf{R}^T(\Delta\alpha, -\xi, \eta) \bar{e}^{\text{LA}}. \quad (15.78)$$

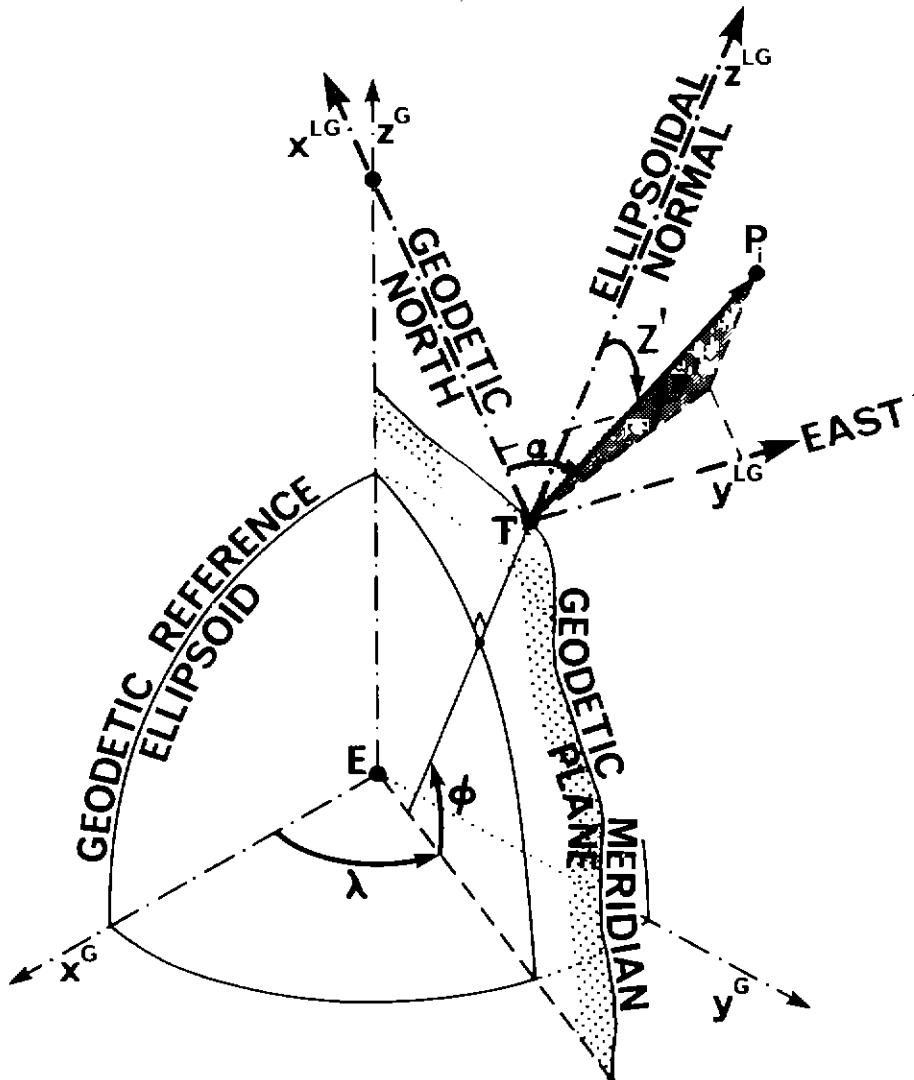


FIG. 15.25. Local geodetic system.

To maintain consistency among the four systems, this equation must be equivalent to eqn. (6), which describes the same transformation only using different transformation parameters. The following equation, hence, must be valid:

$$\begin{aligned} & \mathbf{R}_3(\pi - \Lambda) \mathbf{R}_2\left(\frac{\pi}{2} - \Phi\right) \mathbf{P}_2 \\ &= \mathbf{R}(\epsilon_x, \epsilon_y, \epsilon_z) \mathbf{R}_3(\pi - \lambda) \mathbf{R}_2\left(\frac{\pi}{2} - \phi\right) \mathbf{P}_2 \mathbf{R}^T(\Delta\alpha, -\xi, \eta). \end{aligned} \quad (15.79)$$

In this equation,  $\epsilon_x, \epsilon_y, \epsilon_z, \Delta\alpha, \xi, \eta, \Lambda - \lambda, \Phi - \phi$  would normally be very small quantities. Thus their trigonometric functions can be developed into power series and only the first terms retained. Doing this results in the following condition [VANÍČEK AND CARRERA, 1985]:

$$\begin{bmatrix} \Delta\alpha \\ \xi \\ \eta \end{bmatrix} = \begin{bmatrix} (\Lambda - \lambda)\sin\phi \\ \Phi - \phi \\ (\Lambda - \lambda)\cos\phi \end{bmatrix} - \begin{bmatrix} \cos\phi(\epsilon_x\cos\lambda + \epsilon_y\sin\lambda) + \epsilon_z\sin\phi \\ \epsilon_x\sin\lambda - \epsilon_y\cos\lambda \\ -\sin\phi(\epsilon_x\cos\lambda + \epsilon_y\sin\lambda) + \epsilon_z\cos\phi \end{bmatrix}, \quad (15.80)$$

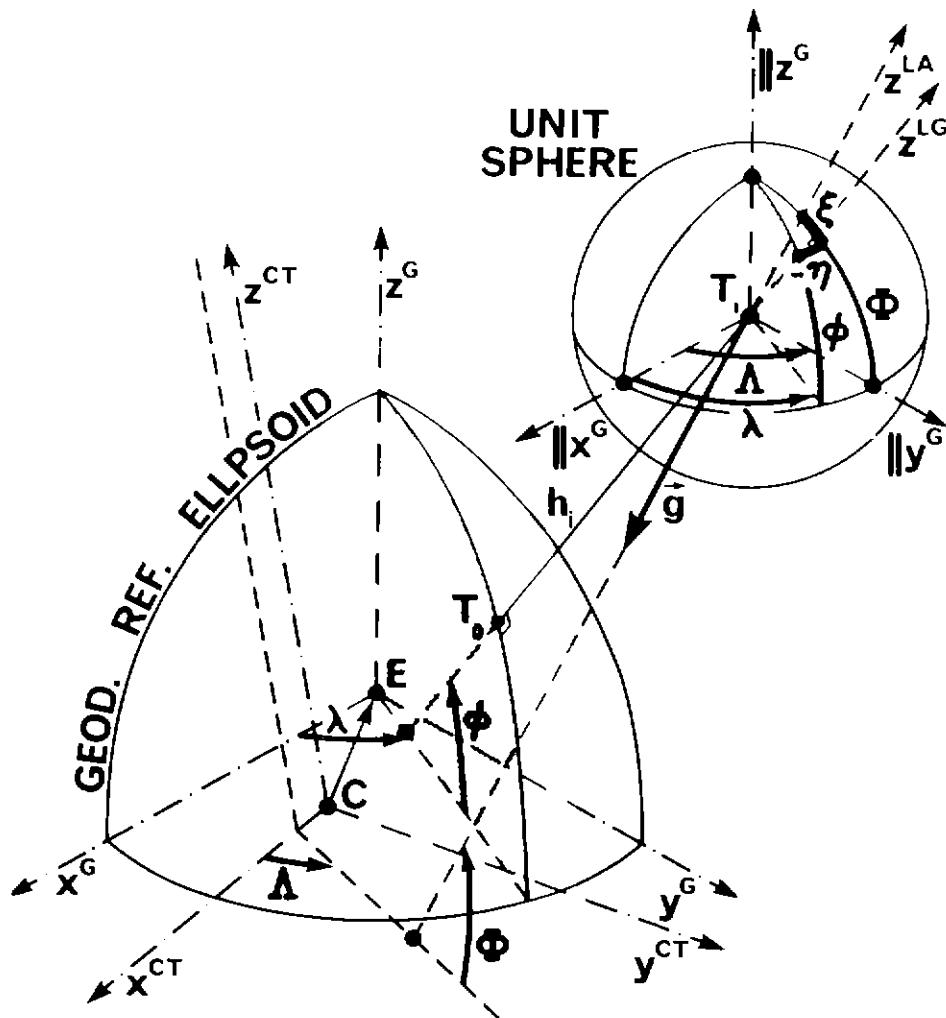


FIG. 15.26. Parallelism conditions.

which must be satisfied by all the quantities involved: geodetic,  $\epsilon_x$ ,  $\epsilon_y$ ,  $\epsilon_z$ ,  $\phi$ ,  $\lambda$ ,  $\alpha$ ; astronomical,  $\Phi$ ,  $\Lambda$ ,  $A$ ; and the surface deflection components,  $\xi$ ,  $\eta$ . It is interesting to note that eqn. (80) can be simplified to

$$\begin{bmatrix} \Delta\alpha \\ \xi \\ -\eta \end{bmatrix} = \begin{bmatrix} (\Lambda - \lambda)\sin\phi \\ \Phi - \phi \\ -(\Lambda - \lambda)\cos\phi \end{bmatrix} - \mathbf{R}_2(\phi - \pi)\mathbf{R}_3(\lambda - \pi) \begin{bmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_z \end{bmatrix}. \quad (15.81)$$

It should be also pointed out that if the G system (and the reference ellipsoid with it) is positioned and oriented with respect to the CT system locally at a point  $T_0$ , a *topocentric set of datum position parameters* is needed. These six independent parameters can be, for instance  $\phi_0$ ,  $\lambda_0$ ,  $\alpha_0$ ,  $\xi_0$ ,  $\eta_0$ ,  $N_0$  [VANIČEK AND WELLS, 1974]. If this mode of datum positioning (and orientation) is used, then eqns. (81) must take a special form, namely:

$$\begin{bmatrix} \Delta\alpha \\ \xi \\ -\eta \end{bmatrix} = \begin{bmatrix} (\Lambda - \lambda)\sin\phi \\ \Phi - \phi \\ -(\Lambda - \lambda)\cos\phi \end{bmatrix} - \mathbf{R}_2(\phi - \pi)\mathbf{R}_3(\lambda - \pi) \begin{bmatrix} \cos\phi_0 \cos\lambda_0 \\ \cos\phi_0 \sin\lambda_0 \\ \sin\lambda_0 \end{bmatrix} \Delta_0, \quad (15.82)$$

where  $\Delta A$  is the misalignment angle reckoned around the ellipsoidal normal at  $T_0$ .

We note that if the G system was properly aligned with the CT system, i.e., if  $\epsilon_x = \epsilon_y = \epsilon_z = 0$ , the second term on the right-hand side of eqns. (81) and (82) would vanish. We would be left with a system of three very simple, but very important, equations:

$$\boxed{\Delta A = A - \alpha = (\Lambda - \lambda) \sin \phi}, \quad (15.83)$$

known as the *Laplace equation* for azimuths, and

$$\boxed{\Phi - \phi = \xi} \quad (15.84)$$

$$\boxed{(\Lambda - \lambda) \cos \phi = \eta}, \quad (15.85)$$

which are the defining equations for the meridian and prime vertical components of the deflection of the vertical (cf. §6.4) in terms of astronomical and geodetic coordinates in the absence of any misalignment. It should be clear that, since the equations are formulated for a point on the earth's surface, the deflections are of the surface type. If the point happens to be on the geoid, then we have the geoidal deflections.

The above three equations constitute the topocentric *conditions for parallelism* for the G and CT systems. If the parallelism is desired, then eqns. (83) to (85) must be satisfied for all points on the earth's surface (including the origin  $T_0$ ), bearing in mind that  $\Phi$ ,  $\Lambda$ ,  $A$  are directly observable.

It is of interest to realize that  $\Delta A$  is the angle between the  $x^{\text{LG}}$ - and  $x^{\text{LA}}$ -axes. Using (85), we can write the Laplace equation in yet another form: namely,

$$A - \alpha = \Delta A = \eta \tan \phi. \quad (15.86)$$

An alternative set of parallelism condition equations can be obtained by considering the observable quantities  $A$  and  $Z$ . Rotating a unit vector in the LA system into the LG system (see FIGS. 3, 25, and 26), we obtain

$$\bar{e}^{\text{LG}} = \mathbf{R}_3(\Delta A) \mathbf{R}_2(-\xi) \mathbf{R}_1(\eta) \bar{e}^{\text{LA}}, \quad (15.87)$$

or

$$\begin{bmatrix} \cos \alpha \sin Z' \\ \sin \alpha \sin Z' \\ \cos Z' \end{bmatrix} \doteq \begin{bmatrix} 1 & \Delta A & \xi \\ -\Delta A & 1 & \eta \\ -\xi & -\eta & 1 \end{bmatrix} \begin{bmatrix} \cos A \sin Z \\ \sin A \sin Z \\ \cos Z \end{bmatrix}.$$

Expanding the left-hand side into a Taylor series at  $(A, Z)$ , we get

$$\begin{aligned} (A - \alpha) \begin{bmatrix} \sin A \sin Z \\ -\cos A \sin Z \\ 0 \end{bmatrix} + (Z - Z') \begin{bmatrix} -\cos A \cos Z \\ -\sin A \cos Z \\ \sin Z \end{bmatrix} &\doteq \\ &\doteq \begin{bmatrix} 0 & \Delta A & \xi \\ -\Delta A & 0 & \eta \\ -\xi & -\eta & 0 \end{bmatrix} \begin{bmatrix} \cos A \sin Z \\ \sin A \sin Z \\ \cos Z \end{bmatrix}. \end{aligned} \quad (15.88)$$

The third equation gives directly

$$Z - Z' = -\xi \cos A - \eta \sin A. \quad (15.89)$$

Multiplication of the first equation by  $\sin A$ , the second by  $\cos A$ , subtraction of the second from the first, and substitution for  $\Delta A$  from eqn. (86) yields:

$$A - (\xi \sin A - \eta \cos A) \cot Z - \alpha = \eta \tan \Phi. \quad (15.90)$$

These equations can be used when precise zenith distances are observed in the network, i.e., in the three-dimensional approach (see §17.1). We note that the second term on the left-hand side of (90) is simply a correction to the observed astronomical azimuth  $A$  to relate it to the same ellipsoidal normal (see FIG. 25) as the geodetic azimuth  $\alpha$ . We will return to these two sets of conditions for parallelism in the context of three-dimensional (see §17.1) and horizontal (§18.1) networks.

We can now finally formulate the transformation of the natural astronomical (physically meaningful) quantities  $(\Phi, \Lambda, A, Z, H)$  into geodetic (conventional) quantities  $(\phi, \lambda, \alpha, Z', h)$ . If the parallelism conditions are satisfied, the transformations  $\phi \leftrightarrow \Phi$  and  $\lambda \leftrightarrow \Lambda$  are given by (84) and (85) and the transformations  $\alpha \leftrightarrow A$  and  $Z' \leftrightarrow Z$  by (90) and (89). The accuracy of  $\phi, \lambda, \alpha$ , and  $Z'$  so obtained is dictated by the accuracy of  $\Phi, \Lambda, A$ , and  $Z$  (about  $0.1''$  to  $0.2''$  as seen in §15.2) and by the accuracy of the known values of  $\xi$  and  $\eta$  (about  $1''$ , see §24.3). The inverse transformations are hardly ever used. It should be noted that the geodetic coordinates so derived refer to the same reference ellipsoid (the same position within the earth as well as the same shape) as the one used for  $\xi$  and  $\eta$ . Also, it is clear that if the reference ellipsoid is not aligned to the CT system, in which  $\Phi$  and  $\Lambda$  are given, then the misalignment angles must be taken into account, as should be clear from the foregoing explanations.

The height above the ellipsoid is related to the height above the sea level  $H$  and geoidal height  $N$  simply through (7.3). The accuracy of the transformations  $h \leftrightarrow H$  is limited by the accuracy of  $N$ , presently about 1 m at best, and that of  $h$  (cf. §16.1 and §17.1) and  $H$  (cf. §16.4 and Chapter 19).

(d) The next transformation to be considered is that between coordinates referred to two different datums. In this transformation it is necessary to account for the difference in the location of the geometrical centre of each reference ellipsoid, the difference in size and shape of the two ellipsoids, and the difference in orientation.

Consider the ellipsoids with sizes and shapes defined by  $(a_1, f_1)$  and  $(a_2, f_2)$ —or alternatively  $(a_1, b_1)$  and  $(a_2, b_2)$ —and the locations of their geometrical centres with respect to the earth's centre of mass defined by  $\vec{r}_E^1$  and  $\vec{r}_E^2$ . Their misalignment angles with respect to the CT system are  $(\epsilon_{x1}, \epsilon_{y1}, \epsilon_{z1})$ ,  $(\epsilon_{x2}, \epsilon_{y2}, \epsilon_{z2})$ , respectively. Let us also denote the coordinates of a point referred to the first datum by  $(\phi_1, \lambda_1, h_1)$ ; we wish to find coordinates  $(\phi_2, \lambda_2, h_2)$  of the same point referred to the second datum.

There are two techniques for obtaining  $(\phi_2, \lambda_2, h_2)$  as functions of  $(\phi_1, \lambda_1, h_1)$ . The first, a direct approach, is to find CT coordinates from (63) and (64), and then

find  $(\phi_2, \lambda_2, h_2)$  using either the iterative method or the closed form inverse solution, as shown earlier under (a) in this section. The second technique, which we show here, is a differential technique that can be applied when the parameter differences  $\delta a = a_2 - a_1$ ,  $\delta f = f_2 - f_1$ ,  $\delta x_E = x_{E2} - x_{E1}, \dots, \delta \epsilon_z = \epsilon_{z2} - \epsilon_{z1}$  for the two datums are sufficiently small.

Let the CT coordinates of a point referred to a geodetic datum be given by eqn. (74). This can be rewritten for small misalignment angles as

$$\begin{aligned}\bar{r}^{CT} &= \begin{bmatrix} x \\ y \\ z \end{bmatrix}^G + \begin{bmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{bmatrix}^G \begin{bmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_z \end{bmatrix} + \begin{bmatrix} x_E \\ y_E \\ z_E \end{bmatrix}^{CT} \\ &= \bar{r}^G + \mathbf{T}^G \bar{\epsilon} + \bar{r}_E^{CT},\end{aligned}\quad (15.91)$$

as the reader can check for himself. An identical equation can be written for the other datum. Distinguishing between the two datums by added subscripts and subtracting the first from the second equation, we get

$$\bar{r}_2^G - \bar{r}_1^G + \mathbf{T}^G (\bar{\epsilon}_2 - \bar{\epsilon}_1) + \bar{r}_{E2}^{CT} - \bar{r}_{E1}^{CT} \doteq \bar{0}. \quad (15.92)$$

Expressing now the geodetic Cartesian coordinates  $\bar{r}_1^G, \bar{r}_2^G$  in terms of corresponding curvilinear geodetic coordinates from eqns. (63), we get, after lengthy development,

$$\bar{r}_2^G - \bar{r}_1^G = \mathbf{J} \begin{bmatrix} \phi_2 - \phi_1 \\ \lambda_2 - \lambda_1 \\ h_2 - h_1 \end{bmatrix} + \mathbf{B} \begin{bmatrix} \delta a \\ \delta f \end{bmatrix}, \quad (15.93)$$

where, using spherical approximations ( $f = 0, N = M = a, h = 0$ ),

$$\mathbf{J} \doteq \begin{bmatrix} -a \sin \phi \cos \lambda & -a \cos \phi \sin \lambda & \cos \phi \cos \lambda \\ -a \sin \phi \sin \lambda & a \cos \phi \cos \lambda & \cos \phi \sin \lambda \\ a \cos \phi & 0 & \sin \phi \end{bmatrix}. \quad (15.94)$$

$$\mathbf{B} \doteq \begin{bmatrix} \cos \phi \cos \lambda & a \sin^2 \phi \cos \phi \cos \lambda \\ \cos \phi \sin \lambda & a \sin^2 \phi \cos \phi \sin \lambda \\ \sin \phi & a(\sin^2 \phi - 2)\sin \phi \end{bmatrix}. \quad (15.95)$$

Substitution of eqn. (93) in eqn. (92) yields

$$\mathbf{J} \left( \begin{bmatrix} \phi_2 \\ \lambda_2 \\ h_2 \end{bmatrix} - \begin{bmatrix} \phi_1 \\ \lambda_1 \\ h_1 \end{bmatrix} \right) + \mathbf{B} \begin{bmatrix} \delta a \\ \delta f \end{bmatrix} + \begin{bmatrix} \delta x_E \\ \delta y_E \\ \delta z_E \end{bmatrix} + \mathbf{T} \begin{bmatrix} \delta \epsilon_x \\ \delta \epsilon_y \\ \delta \epsilon_z \end{bmatrix} \doteq \bar{0}, \quad (15.96)$$

and the desired transformation equation finally reads:

$$\begin{bmatrix} \phi_2 \\ \lambda_2 \\ h_2 \end{bmatrix} \doteq \begin{bmatrix} \phi_1 \\ \lambda_1 \\ h_1 \end{bmatrix} - \mathbf{J}^{-1} \left( \begin{bmatrix} \delta x_E \\ \delta y_E \\ \delta z_E \end{bmatrix} + \mathbf{T} \begin{bmatrix} \delta \epsilon_x \\ \delta \epsilon_y \\ \delta \epsilon_z \end{bmatrix} + \mathbf{B} \begin{bmatrix} \delta a \\ \delta f \end{bmatrix} \right), \quad (15.97)$$

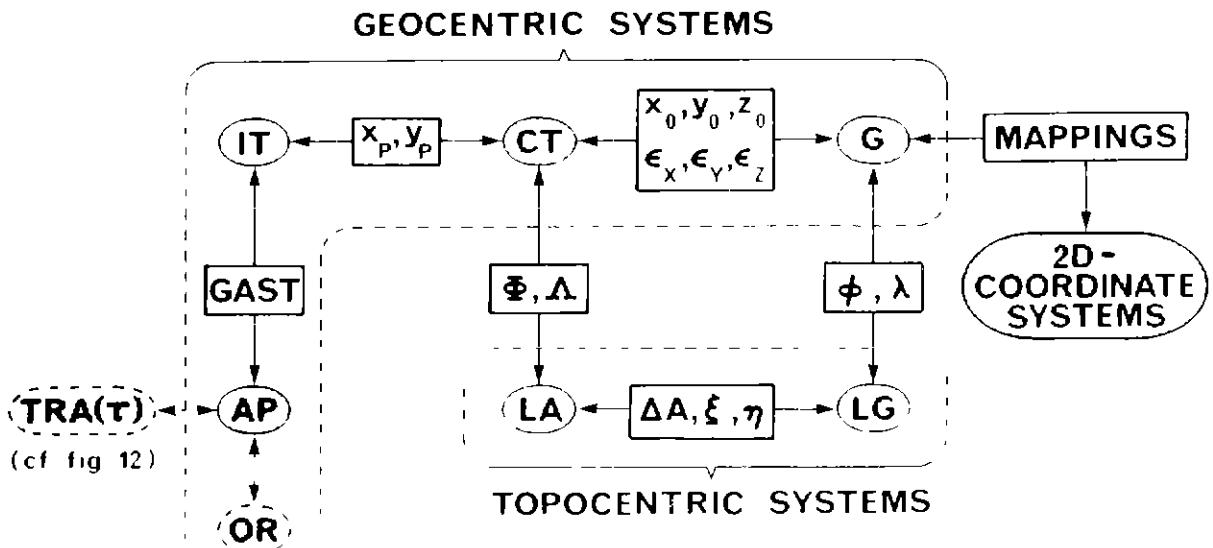


FIG. 15.27. Commutative diagram of transformations of terrestrial positions.

where

$$\mathbf{J}^{-1} = \begin{bmatrix} -\sin \phi \cos \lambda / a & -\sin \phi \sin \lambda / a & \cos \phi / a \\ -\sin \lambda / (a \cos \phi) & \cos \lambda / (a \cos \phi) & 0 \\ \cos \phi \cos \lambda & \cos \phi \sin \lambda & \sin \phi \end{bmatrix}. \quad (15.98)$$

The matrices  $\mathbf{B}$ ,  $\mathbf{J}$ , and  $\mathbf{T}$  can be evaluated on either of the two datums, since the differences in the ellipsoids are assumed to be small. The  $\mathbf{T}$  matrix is best written as

$$\mathbf{T} = a \begin{bmatrix} 0 & -\sin \phi & \cos \phi \sin \lambda \\ \sin \phi & 0 & -\cos \phi \cos \lambda \\ -\cos \phi \sin \lambda & \cos \phi \cos \lambda & 0 \end{bmatrix}. \quad (15.99)$$

(e) The direct and inverse transformations of the geodetic latitude and longitude  $(\phi, \lambda)$  into two-dimensional Cartesian *map coordinates*  $(x, y)^M$ , and vice versa, can be written in terms of mapping equations usually simply called *mappings*:

$$x = x(\phi, \lambda), \quad y = y(\phi, \lambda), \quad (15.100)$$

$$\phi = \phi(x, y), \quad \lambda = \lambda(x, y). \quad (15.101)$$

As the general theory of mapping lies within the realm of mathematical cartography, mappings will not be dealt with in this book. Suffice it to say here that a multitude of mappings exist, which can be used to represent the horizontal geodetic coordinates, and the interested reader is referred to the literature (e.g., HOTINE [1946, 1947], RICHARDUS AND ADLER [1972], MALING [1973]). Of particular importance to geodesy are conformal mappings that are used in various geodetic computations; these will be treated, more appropriately, in §16.3.

In closing, a commutative diagram showing all the transformations treated in this section is presented in FIG. 27. It is instructive to note that the AP system is the link between the terrestrial and the other systems. The symmetry of the (CT, LA) and (G, LG) pairs, mentioned earlier in this chapter, is also clearly visible. Finally we note the fact that the transformations between these four systems make a closed diagram, which explains why a datum (G system) can be positioned, as shown above under (c), in two different ways: globally and locally.

## CHAPTER 16

### RELATIVE POSITIONING

Relative positioning is the determination of the location of one point with respect to another, either by measuring directly between the two points or by measuring indirectly from the two points to extraterrestrial objects. Further, the type of observations collected and the kind of coordinates desired dictate whether the mathematical model is formulated in a three-dimensional, two-dimensional, or one-dimensional space. The relative three-dimensional positioning, terrestrial and extraterrestrial, is treated in the first section. Relative two-dimensional, horizontal positioning on the reference ellipsoid and the conformal mapping plane are discussed in the second and third sections respectively. Relative one-dimensional, vertical positioning is dealt with in the fourth section.

The problem inverse to relative positioning may be stated as follows: Given the coordinates of two points, compute the direction and distance between these two points. The inverse problems, in three and two dimensions, are treated after the direct problems in the first three sections.

#### 16.1. Relative three-dimensional positioning

The direct problem of *terrestrial, relative three-dimensional positioning* reads as follows: Given the coordinates  $(x_i, y_i, z_i)^{CT}$  and the astronomical coordinates  $(\Phi_i, \Lambda_i)$  or, equivalently, the deflection components  $\xi_i, \eta_i$ , of an observation point  $P_i$ , along with the observations of the astronomical azimuth  $(A_{ij})$ , the vertical angle  $(\nu_{ij})$  or the zenith distance  $(Z_{ij})$ , and the spatial distance  $(\rho_{ij})$  to an observed point  $P_j$ , compute the CT coordinates of  $P_j$ . Before attempting the solution, some comments on the quantities involved are in order.

To begin with, it is assumed that the given astronomical quantities  $(\Phi_i, \Lambda_i, A_{ij})$  have been corrected for the effect of polar motion so that they refer to the conventional spin axis of the earth (CIO)—cf. §15.1. Let us mention here that the astronomical azimuth  $A$  may be obtained either from astronomical observations (cf. §15.2) or from observations with a gyro-theodolite. While the former technique gives an accuracy of up to  $\sigma_A = 0.4''$  [MUELLER, 1969], the most recently achieved accuracy of gyroscopically determined  $A$  is about  $\sigma_A = 1''$  [GREGERSON, 1980]. Vertical angles, or zenith distances, can be measured with a geodetic theodolite to a theoretical accuracy of about  $\sigma_\nu = 2''$  [RAMSAYER, 1971], if a correction is made for the vertical

refraction as explained in §15.2. Usually, though, there remains a systematic residual of vertical refraction so that the actually achieved accuracy is lower. This systematic component may be eliminated, to a high degree, if two-wavelength instrumentation, employing two-colour light, is used [HUGGETT AND SLATER, 1978]. Another way to eliminate the residual refraction is to solve for it mathematically; this can only be done within a specially designed network of points (see §17.2). Spatial distances, normally determined by an EDM device, also need to be corrected for instrumental errors and refraction using (15.39) or one of the many other existing approximate formulae; the reader interested in more details is referred to, e.g., SAASTAMOINEN [1967]. Horizontal angles or directions are measured with a theodolite and their accuracy can be better than  $1''$  [BOMFORD, 1971].

The solution is obtained by transforming a vector from the LA system to the CT system in three steps:

(a) The observed topocentric position vector, also called the *interstation vector* of  $P_j$  with respect to  $P_i$ , is first formulated as (see FIG. 15.3 and eqn. (15.4))

$$\bar{r}_{ij}^{\text{LA}} = \Delta r_{ij} \bar{e}_{ij}^{\text{LA}} = \Delta r_{ij} \begin{bmatrix} \cos \nu_{ij} \cos A_{ij} \\ \cos \nu_{ij} \sin A_{ij} \\ \sin \nu_{ij} \end{bmatrix}. \quad (16.1)$$

It gives the relative position of  $P_j$  with respect to  $P_i$  in the LA system of  $P_i \equiv T$ .

(b) Then  $\bar{r}_{ij}^{\text{LA}}$  is transformed to the CT system by (15.6) using the given  $\Phi_i$ ,  $\Lambda_i$ .

(c) The position vector of  $P_j$  in the CT system can now be computed by the simple addition of vectors, namely,

$$\bar{r}_j^{\text{CT}} = \bar{r}_i^{\text{CT}} + \Delta \bar{r}_{ij}^{\text{CT}} = \bar{r}_i^{\text{CT}} + \mathbf{R}_3(\pi - \Lambda_i) \mathbf{R}_2\left(\frac{1}{2}\pi - \Phi_i\right) \mathbf{P}_2 \bar{r}_{ij}^{\text{LA}}, \quad (16.2)$$

where  $\bar{r}_i^{\text{CT}} = (x_i, y_i, z_i)^{\text{CT}}$  is given. In §15.4 it was shown how  $\bar{r}_j^{\text{CT}}$  is then transformed into  $\bar{r}_j^G$  and further, if desired, into the curvilinear geodetic coordinates  $(\phi_i, \lambda_i, h_i)$ .

If  $\xi_i, \eta_i$  are given instead of  $\Phi_i, \Lambda_i$ , then the solution is again obtained by transformation from the LA to the CT system, but this time by means of the LG and CT G systems. This procedure can only be used if the misalignment of the G and CT systems is known; here we shall assume that the systems are parallel (cf. §15.4). The solution can be broken down into the following steps:

(a) The observed topocentric position vector  $\bar{r}_{ij}^{\text{LA}}$  is again defined by (1).

(b) It is then transformed into the LG system using (15.87), where  $\xi_i$  and  $\eta_i$  are the components of the surface deflection of the vertical (see §6.4), and  $\Delta A_{ij} = A_{ij} - \alpha_{ij}$  is given as a function of  $\eta_i$  through the Laplace equation (15.83).

(c) The topocentric position vector is further transformed into the G system through (cf. (15.6))

$$\Delta \bar{r}_{ij}^G = \mathbf{R}_3(\pi - \lambda_i) \mathbf{R}_2\left(\frac{1}{2}\pi - \phi_i\right) \mathbf{P}_2 \bar{r}_{ij}^{\text{LG}}. \quad (16.3)$$

Clearly,  $\phi_i$  and  $\lambda_i$  must be the same coordinates of  $P_i$  that were used in defining  $\xi_i$  and  $\eta_i$ .

(d) The position vector of  $P$ , in the geodetic system is obtained, again, simply by the addition of vectors:

$$\bar{r}_j^G = \bar{r}_i^G + \Delta \bar{r}_{ij}^G, \quad (16.4)$$

where  $\bar{r}_i^G$  must be known.

(e) Finally, if desired, the transformation of the above into the CT system (see (15.74)) or further into curvilinear geodetic coordinates is carried out.

*Inertial positioning* is another way of obtaining relative three-dimensional positions; it is based on utilizing the inertia of a mass to measure accelerations. The measurements are performed by sensors called *accelerometers*. As we know, an acceleration  $\bar{a}$  of the mass is a result of a force acting on the mass. These three physical quantities are related through the second Newton law, which is the foundation of the mathematical model for inertial positioning.

An accelerometer senses only the component of acceleration that coincides in direction with its longitudinal axis; thus, to obtain the acceleration vector  $\bar{a}$ , three such components ( $a_x, a_y, a_z$ ) have to be measured (FIG. 1). Obeying Einstein's principle of equivalence, a three-dimensional accelerometer standing still on the surface of the earth will register an acceleration of its instrument frame of approximately 981 Gal (cf. §6.1) in the direction pointing toward the local zenith. This is a consequence of the action of the force of gravity on the masses of the accelerometers; the readout will show the reaction of the masses interpreted as a movement of the instrument frame in response to this force. This behaviour is described by the third Newton law: To every force there is an opposite and equal reaction.

An accelerometer mounted in a moving ground vehicle senses the sum of gravity ( $\bar{g}$ ) and the acceleration ( $\bar{a}$ ) of the vehicle with respect to the gravity field, i.e., with

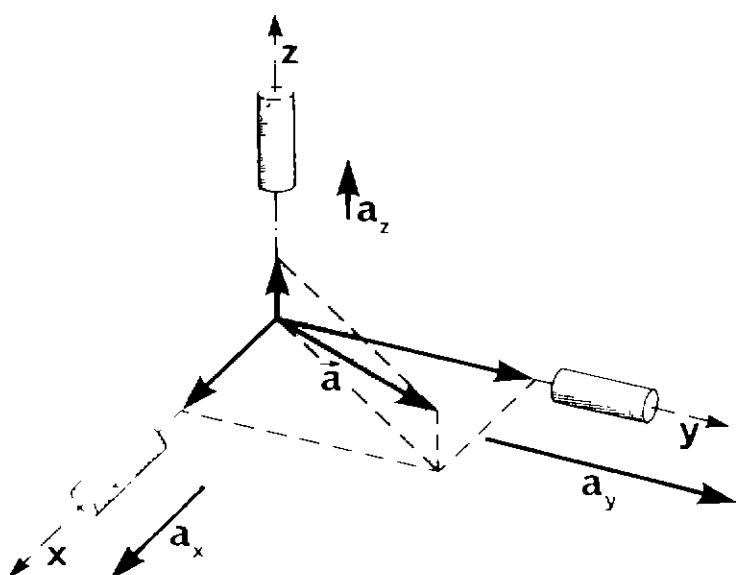


FIG. 16.1 Three-component accelerometer

respect to the earth (cf. FIG. 2). As well, it senses solar and lunar tidal acceleration (cf. §8.1), acceleration due to the sea tide (see §25.3), and that due to polar motion (see §25.4). If the accelerometer moves around, then Coriolis's acceleration (see (9.17)) also affects its response. The former three accelerations are comparatively small and, in the first approximation, are disregarded [KAYTON, 1960]. For positioning, only the acceleration of the vehicle with respect to the earth is needed. This means that the gravity vector  $\bar{g}$ , as a function of the location of the vehicle and Coriolis's acceleration vector, have to be subtracted from the observed total acceleration vector appearing on the output of the accelerometer. The gravity vector has to be known; its determination is addressed in Part V and will not be enlarged upon here.

Once the gravity and Coriolis's acceleration vectors have been subtracted, then the remaining vehicle acceleration vector changes with time, if the vehicle moves, and thus is written as  $\bar{a}(\tau)$ . The relation of this observable  $\bar{a}(\tau)$  to the unknown position vector  $\bar{r}(\tau)$  of the vehicle can then be formulated. First the velocity  $\bar{v}(\tau)$  of the vehicle is related to  $\bar{r}(\tau)$  by

$$\bar{v}(\tau) = \dot{\bar{r}}(\tau) = \frac{d\bar{r}(\tau)}{d\tau}, \quad (16.5)$$

and, inversely,

$$\bar{r}(\tau_1) = \bar{r}(\tau_0) + \int_{\tau_0}^{\tau_1} \bar{v}(\tau) d\tau, \quad (16.6)$$

where  $\bar{r}(\tau_0)$  is the initial position at time  $\tau_0$ . Secondly, the acceleration is related to

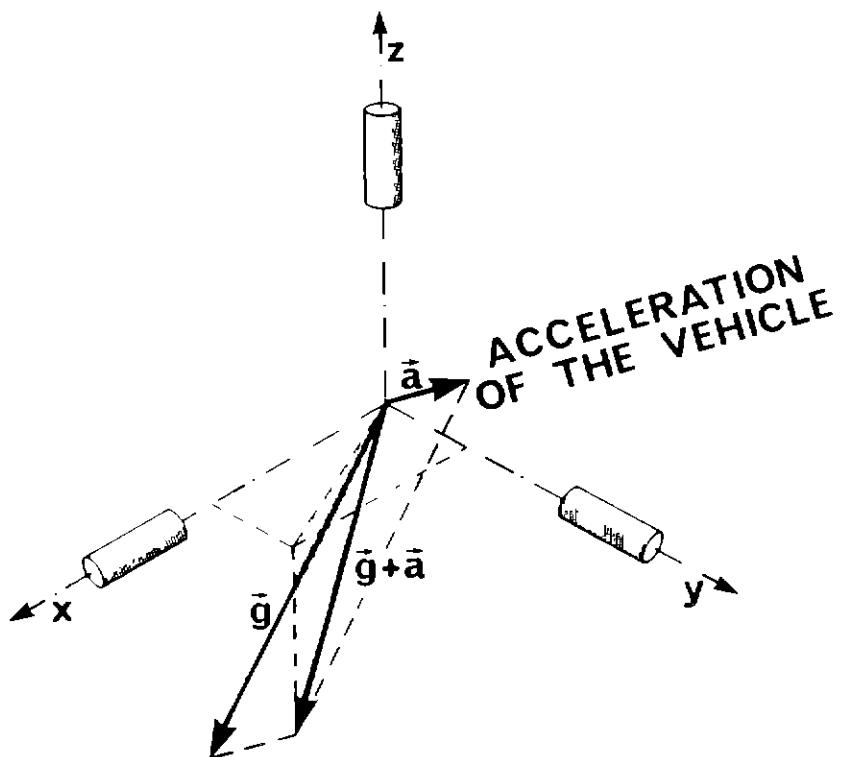


FIG. 16.2. Acceleration of the vehicle and gravity.

the velocity in an analogous fashion by

$$\bar{a}(\tau) = \frac{d\bar{v}(\tau)}{d\tau} = \ddot{\bar{r}}(\tau) = \frac{d^2\bar{r}(\tau)}{d\tau^2}, \quad (16.7)$$

and, inversely,

$$\bar{r}(\tau_1) = \bar{r}(\tau_0) + \bar{v}(\tau_0)(\tau_1 - \tau_0) + \int \int_{\tau_0}^{\tau_1} \bar{a}(\tau) d\tau d\tau. \quad (16.8)$$

If the initial velocity  $\bar{v}(\tau_0)$  is zero, i.e., if the vehicle starts to move from a stationary position at time  $\tau_0$ , the above simplifies to

$$\bar{r}(\tau_1) = \bar{r}(\tau_0) + \int \int_{\tau_0}^{\tau_1} \bar{a}(\tau) d\tau d\tau. \quad (16.9)$$

This is the basic mathematical model for inertial positioning. It is explicit in unknown parameters (cf. §10.2).

The assumption made above that  $\bar{a}(\tau)$ , obtained from the three-component accelerometer fastened to the vehicle, is always measured in the same coordinate system is false. The instrument frame at an instant  $\tau$  is generally not parallel to that at the instant  $\tau_0$ , because the vehicle pitches, rolls, and yaws as it moves along. The monitoring of this motion can be performed by, for instance, three free gyroscopes. Theoretically, if the gyroscope's spin axis is free to move, then it maintains the same fixed direction with respect to the inertial space (cf. §15.1) as long as the flywheel keeps spinning [SCARBOROUGH, 1958]. Three free gyroscopes, with non-parallel spin axes, can thus provide a fixed directional reference for the moving instrument frame. All that has to be done then is to monitor the varying misalignment of the instrument frame with respect to this fixed directional reference system. The spatial misalignment is uniquely depicted by three independent angles describing, for instance, the rotations around the orthogonal axes  $x$ ,  $y$ ,  $z$  (cf. §3.3) of the instrument frame, as shown in FIG. 3.

It is easy to see that, if the rotations are recorded (the same way the components of  $\bar{a}(\tau)$  are recorded) as they vary with time, the varying direction of  $\bar{a}(\tau)$  with respect to the inertial frame can be reconstructed for every instant  $\tau$  through three time-varying rotation matrices. Thus  $\bar{a}(\tau)$ , and the last term (the double integral) in (9) as well, can be referred to a fixed coordinate system throughout the measurement process. Most commonly, the system is kept physically aligned to the LA system while the output  $\bar{a}(\tau)$  is mathematically transformed to a selected G system by modelling the angular changes between the LA system, from point to point, and the G system. The modelling involves the simulation of the vehicle's motion with respect to the desired reference ellipsoid and to the earth's gravity field. The position vector of the moving vehicle (point) in, say, the G system is defined as

$$\bar{r}^G(\tau) = \bar{r}^G(\tau_0) + \Delta \bar{r}^G(\tau), \quad (16.10)$$

**THIS FRAME HELD FIXED  
IN SPACE BY GYROSCOPES**

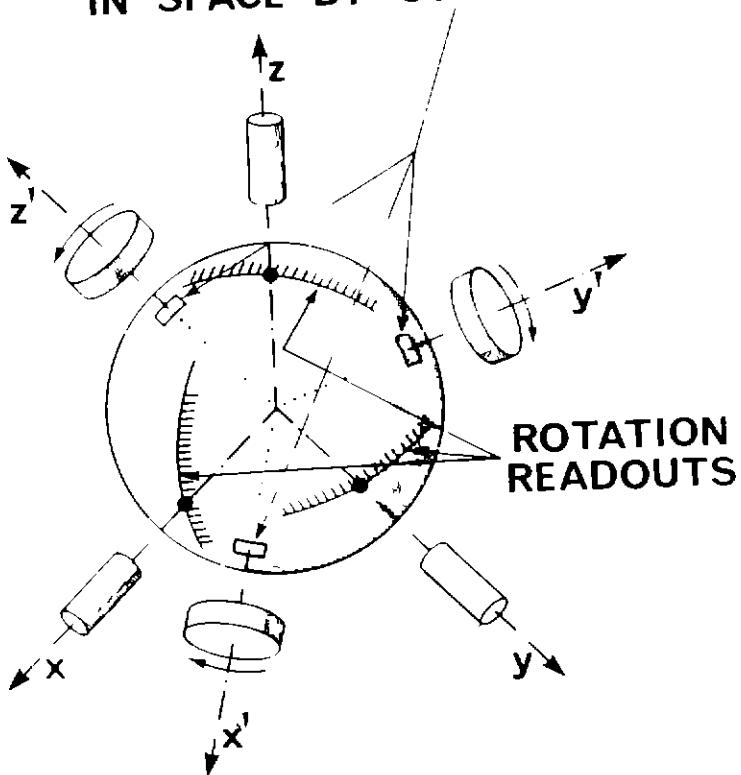


FIG. 16.3. Rotation of instrument frame.

where  $\bar{r}^G(\tau_0)$  is the position vector of the starting point, and  $\Delta\bar{r}^G(\tau)$  is the double integral from (9).

In reality, the above described model must be expanded to include various nuisance parameters: three biases of the component accelerometers and the biases of the gyroscopes, all with respect to the instrument frame, plus those that model other instrument imperfections responsible for the drift of the system, bringing the total number of unknowns to as many as 27. How then can we solve for all these unknowns from only three explicit equations (9)? Additional equations that describe the variation of different nuisance parameters with time [ADAMS, 1977] can be added to the equations describing the needed parameters of the gravity field. The resultant system of equations has the general form of the Kalman filter developed in §14.6. More than one point position has to be known to calibrate the system (eliminate some nuisance parameters) and the system is thus used only as an interpolation tool between known points. Also frequent stops—zero velocity updates called ZUPTs—are required to eliminate still more parameters. Prior to 1975, accuracies in relative position determination of coordinate differences of about  $\sigma = 1$  or 2 m were achieved over distances of about 50 km [GREGERSON, 1975]. Nowadays, following special precautions in the field operations, the accuracy is one order of magnitude better. For more details, interested readers are referred to BRITTING [1971] and DRAPEK [1977].

Let us now turn to the direct problem of relative positioning by extraterrestrial methods. Common to the existing methods is the fact that simultaneous measurements are made from the two points in question to one or more space objects. Depending on the method employed, it may be possible to obtain only the direction of the interstation vector (direction cosines) or we may be able to get the complete vector (coordinate differences).

*Relative positioning by directions to satellites* is basically very simple. Simultaneous direction measurements from two tracking stations  $P_i, P_j$  to the first satellite position  $S_1$  yield two unit vectors,  $\bar{e}_i^1$  and  $\bar{e}_j^1$ , which lie in the plane  $P_i P_j S_1$  (see FIG. 4). From (15.44), one readily gets

$$\bar{e}_i^{1CT} = \mathbf{R}_2(-x_p) \mathbf{R}_1(-y_p) \mathbf{R}_3(\text{GAST}) \begin{bmatrix} \cos \delta_1 \cos \alpha_1 \\ \cos \delta_1 \sin \alpha_1 \\ \sin \delta_1 \end{bmatrix}, \quad (16.11)$$

where  $x_p, y_p$ , and GAST have been defined in §15.1, and  $\alpha$  and  $\delta$  are the apparent places determined by photographing the position  $S_1$  of the satellite against the star background from tracking station  $P_i$ . The vector  $\bar{e}_j^1$  is defined in a similar fashion. The vector product of the two unit vectors defines the normal to the plane; namely,

$$\bar{n}_1 = \bar{e}_i^1 \times \bar{e}_j^1. \quad (16.12)$$

Similarly, the second plane is defined by its unit normal vector  $\bar{n}_2 = \bar{e}_i^2 \times \bar{e}_j^2$ . The unit interstation vector must be orthogonal to both normal vectors and is thus given

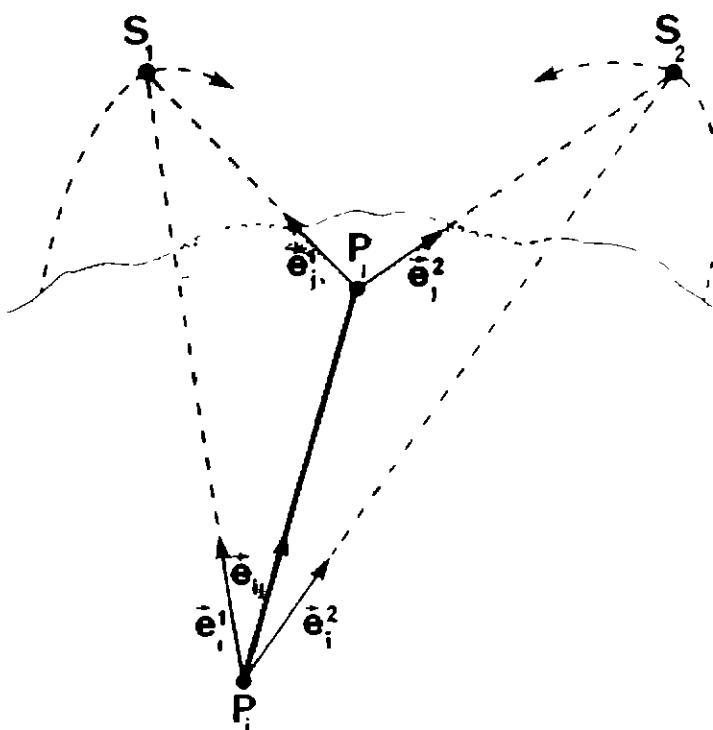


FIG. 16.4. Relative positioning by directions to satellites.

by

$$\bar{e}_{ij} = \bar{n}_1 \times \bar{n}_2 = (\bar{e}_i^1 \times \bar{e}_j^1) \times (\bar{e}_i^2 \times \bar{e}_j^2). \quad (16.13)$$

If simultaneous observations to more than two satellite positions are available, then a system of equations of this kind can be written for  $\bar{e}_{ij}$ .

A vector  $\bar{e}_{ij}$  derived in this manner is obviously expressed in the CT system. To obtain the astronomical azimuth and vertical angle of  $\bar{e}_{ij}$ , we simply rotate  $\bar{e}_{ij}$  into the LA system using equations inverse to (15.6), and resolve its components into  $A_{ij}$  and  $v_{ij}$  using (15.5). The length  $\Delta r_{ij}$  of the interstation vector cannot be obtained from the direction observations alone. The SAO, in their earlier work, used this approach. It routinely yielded accuracies of about  $\sigma = 1''$  in the direction of  $\bar{e}_{ij}$  between terrestrial points [AARDOOM ET AL., 1967]. This corresponds to positional uncertainties of about 5 m for exactly known distances  $\Delta r_{ij}$  of the order of 1000 kilometres.

Next, let us consider *relative positioning by (satellite) range differences*. This approach, combined with the TRANSIT system, has become known as the *translocation technique* [WESTERFIELD AND WORSLEY, 1966]. Again, at least two tracking stations must track a common satellite, i.e., make a sequence of simultaneous measurements (FIG. 5). Usually more than four such measurements are made at each station. The reason for the simultaneity is to get the systematic errors, that affect both tracking stations in a similar fashion, to cancel when the equations for the interstation vector  $\Delta \bar{r}_{ij}$  are formed. The main sources of systematic errors are the

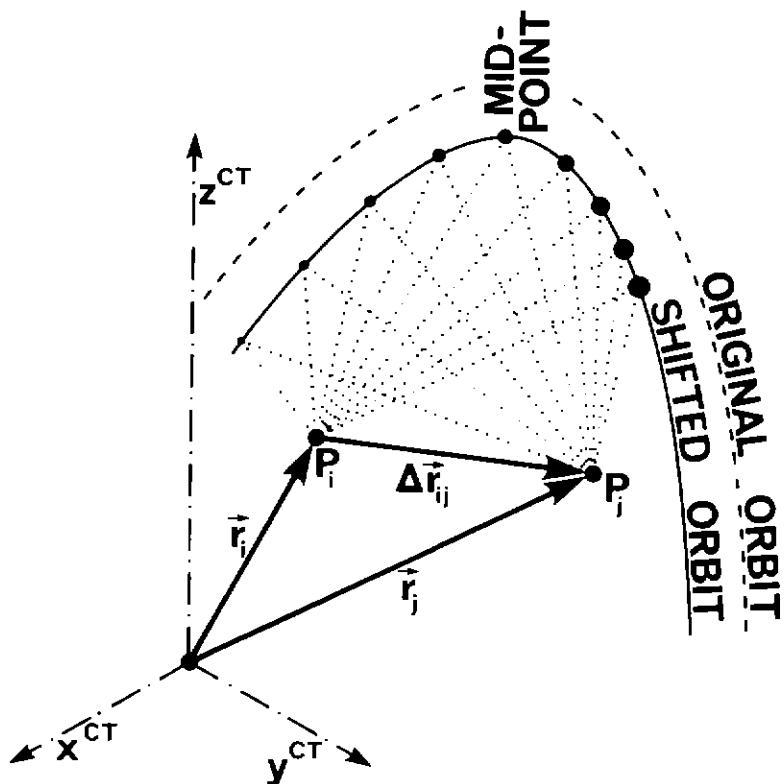


FIG. 16.5. Translocation technique.

satellite orbit and the inadequately modelled tropospheric and ionospheric delays. Since refraction is thought to behave, in the first approximation, in a fashion symmetric with respect to the zenith, the range differences are often chosen to be symmetric to the mid-point of the satellite pass (see FIG. 5) to ensure the cancellation of delay errors. Using the collected data at each station  $P_i$  and  $P_j$ , the position vectors  $\bar{r}_i$  and  $r_j$  are computed using the point positioning mathematical model (15.52) twice once for  $P_i$  and a second time for  $P_j$ . Then  $\Delta\bar{r}_{ij} = \bar{r}_j - \bar{r}_i$  is formed; it is referred to the CT system, since the position vectors are in the CT system.

The translocation technique has been further refined by KOUBA AND WELLS [1976]; in their approach, each orbit is allowed to shift in a parallel manner into a new position such that the interstation vector, as determined from the whole family of orbits, best satisfies all the observations. This technique has become known as the *semidynamic mode of translocation* and forms the basis of the computer program GEODOP. It is readily extended to include several pairs of tracking stations thereby allowing a whole network to be adjusted simultaneously (see §17.3). The technique of translocation in the semidynamic mode based on TRANSIT has achieved an accuracy level of about  $\sigma = 0.4$  m in all three components of the interstation vector [WELLS ET AL., 1976; KOUBA, 1980]. Let us mention in passing that the integrated Doppler count observations obtainable from the GPS allow the same technique to be used with NAVSTAR. The results, however, are much less accurate than those from GPS differential ranging described below.

The translocation concept has also been applied to *relative positioning by ranging to satellites* using laser [LATIMER AND CAPOSHKIN, 1977]. Accuracies better than 1 m have been obtained for the length of the interstation vector. There is nothing to prevent one from using ranging to the moon in the same mode.

Of particular importance is relative positioning by GPS, which is known as *differential GPS positioning*. Because NAVSTAR is a more flexible system than TRANSIT, due to the fact that several satellites are always visible at any one time, there is a greater variety of different observables available with the GPS. In addition to the above mentioned range differences that can be used in a translocation mode, simultaneously observed P-code or C/A-code ranges can be used to create *differential ranges* defined thus:

$$\Delta\rho_{ij}^k = \rho_j^k - \rho_i^k. \quad (16.14)$$

We can also use the carrier signals, however, to create differential ranges. These can be measured inherently much more accurately than differential ranges from the codes. And, what may be a very advantageous trait, these can be measured without any knowledge of either code by stripping the carrier of all the modulation, e.g., through squaring it. The price one pays for this is that the basic timing capability is lost, and the differential ranges become ambiguous; while we may know the value of the differential range to 1 to 2 mm, we are uncertain about the integral multiple of half-wavelengths of the carrier (i.e., either 12.2 cm, or 9.5 cm respectively for the two carriers) contained in the differential range. This problem is, however, surmountable (COUNSELMAN AND GOUREVITCH, 1981).

The *relative positioning by differential range mathematical model* is quite simple; it

reads [BOSSLER ET AL., 1980; VANÍČEK ET AL., 1984]:

$$\left( \bar{e}_i^k + \bar{e}_j^k \right) \Delta \bar{r}_{ij} = - \left( 1 + \bar{e}_i^k \bar{e}_j^k \right) \Delta \rho_{ij}^k = - \left( 2 - \frac{(\omega^k)^2}{2} \right) \Delta \rho_{ij}^k, \quad (16.15)$$

where  $\omega^k$  is the (paralactic) angle at  $S_k$  subtended by  $\Delta \bar{r}_{ij}$ . Here, the differential ranges are assumed to have been corrected for the differences in atmospheric delays. This correction is much easier to model than the whole delay correction. Also, models have been designed for differential range differences, differences of range differences, etc. [GOAD AND REMONDI, 1984], which further reduce both refraction and orbit errors at the cost of somewhat weakening the geometry of the configurations. In the very near future, relative accuracy of this positioning mode is likely to reach  $10^{-7}$  for observing sessions lasting an hour or less [LANGLEY ET AL., 1984].

Another method of relative three-dimensional positioning that should be discussed here is *astronomical radio-interferometry*, also known as long or very long base line interferometry (LBI or VLBI) [BROTON ET AL., 1967]. The method uses signals (emitted by extragalactical bodies, called quasars) that are periodic over a wide range of radio frequencies from a few MHz to several GHz. The quasar signal is many times weaker than satellite signals, thus more complex receivers and much larger directional antennas (dishes) must be used to receive it. Quasars, being far away, are virtually dimensionless in the sky and thus particularly suited as reference points.

The principle of the LBI or VLBI method is illustrated in FIG. 6. The difference, called *time delay*  $\tau$ , in the arrival time of the same wave-front at two observatories, more precisely at the radio centres  $P_i, P_j$  of the two dishes, is used to compute the projected length of the base line onto the direction to the quasar. Given the unit vector  $\bar{e}_s$  in the direction of the source and the interstation vector  $\Delta \bar{r}_{ij}$ , the scalar product yields

$$\bar{e}_s \cdot \Delta \bar{r}_{ij} = |\bar{e}| |\Delta \bar{r}_{ij}| \cos \psi = \tau c, \quad (16.16)$$

where  $\psi$  is the spatial angle between the base line and the direction to the source, and  $c$  is the velocity of light. The mathematical model is then

$$\tau = \frac{1}{c} \bar{e}_s \cdot \Delta \bar{r}_{ij}. \quad (16.17)$$

The vector  $\bar{e}_s$  is considered to be in the CT system and is given by (11). Clearly, one observed time delay yields one observation equation (17). To solve for the three unknown components of  $\Delta \bar{r}_{ij}$ , at least three time delays to different quasars need be measured. To improve the accuracy of  $\Delta \bar{r}_{ij}$ , more measurements are usually taken and different quasars are used. Also, the Doppler shift of the quasar signal, called the *fringe frequency* here, can be measured and combined with the time delays. Distances of several thousands of kilometres can now be measured to an accuracy of about  $\sigma = 2$  cm [NASA, 1984].

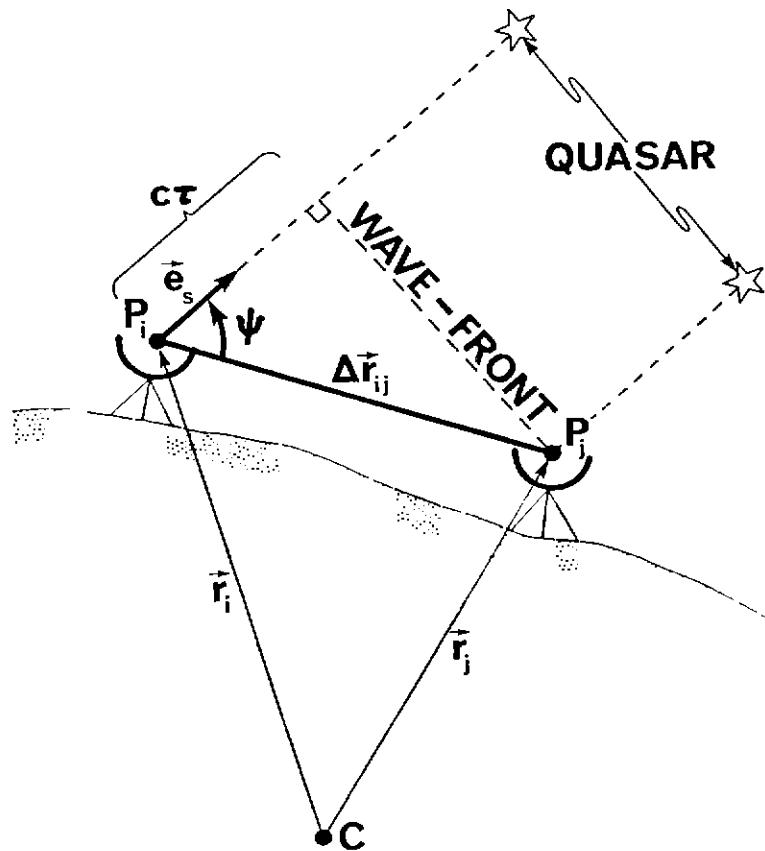


FIG. 16.6. Relative positioning by radio-astronomical interferometry.

It should be pointed out that the difference in the techniques of LBI and VLBI lies in the mode of time synchronization: LBI uses direct connection (cable) while in VLBI, received signals are modulated and the time synchronization is achieved through a post-mission correlation process. Preprocessing and correlating signals from different observatories are elaborate and expensive tasks. Radio-astronomy techniques also place a high demand on instrumentation; nevertheless, mobile systems capable of achieving accuracies at the level of a few centimetres [MACDORAN ET AL., 1978] are available. For the time being, however, because of the dimensions and high costs of the necessary equipment, quasar tracking remains beyond the reach of most surveyors.

Having discussed the direct problem of relative three-dimensional positioning using terrestrial as well as extraterrestrial methods, we can now turn to the *inverse problem of relative three-dimensional positioning*. Given the G coordinates of two points  $P_i$  and  $P_j$ , we are to compute the spatial distance, azimuth, and vertical angle (or a zenith distance) of these two points. The solution is obtained by first evaluating the interstation vector  $\Delta \vec{r}_{ij}^G$ . This vector is then transformed to the LG system of  $P_i$  by means of inverted (3). Finally, the desired quantities are calculated from the following equations:

$$\Delta r_{ij} = |\Delta \vec{r}_{ij}| = (\Delta x_{ij}^2 + \Delta y_{ij}^2 + \Delta z_{ij}^2)^{1/2}, \quad (16.18)$$

$$\nu'_{ij} = \frac{1}{2}\pi - Z'_{ij} = \arcsin(z_{ij}^{\text{LG}}/\Delta r_{ij}), \quad (16.19)$$

and  $\alpha_{ij}$  is obtained from an equation identical with (15.5) except that LG rather than LA coordinates are used. If astronomical, rather than geodetic, quantities are required, then the interstation vector is first transformed to the LA system (see (15.6) and (15.87)) as

$$\begin{aligned} \bar{r}_{ij}^{\text{LA}} &= \mathbf{P}_2 \mathbf{R}_2 (\Phi_i - \frac{1}{2}\pi) \mathbf{R}_3 (\Lambda_i - \pi) \Delta r_{ij}^{\text{CT}} \\ &= \mathbf{R}_1 (-\eta_i) \mathbf{R}_2 (\xi_i) \mathbf{R}_3 (-\Delta A_{ij}) \bar{r}_{ij}^{\text{LG}}. \end{aligned} \quad (16.20)$$

The astronomical azimuth and vertical angle are then given by (15.5), while the distance is computed from (18). If the azimuth and vertical angle from  $P_j$  to  $P_i$  are desired, then the subscripts  $i$  and  $j$  in all the above equations are reversed.

To this point, we have purposely avoided discussing the formal way in which the accuracy of positions may be assessed; now we can do it properly. The triplet of estimated coordinates  $\hat{r}_j$  (denoted here simply by  $\hat{x}$ ) of the sought point  $P_j$  has a covariance matrix  $C_{\hat{x}}$  that emerges as part of the solution simply by applying the least-squares method (cf. Chapter 12). As we have already seen ((13.35) and subsequent equations), the multivariate probability density function for  $x$  contains the quadratic form  $(x - \hat{x})^T C_{\hat{x}}^{-1} (x - \hat{x})$  whose value dictates the value of the probability enclosed within the hypersurface of the probability density function. In our case, when  $x$  has only three components, the equation,

$$(x - \hat{x})^T C_{\hat{x}}^{-1} (x - \hat{x}) = C_{\alpha}^2, \quad (16.21)$$

can be interpreted as describing a triaxial ellipsoid, with the centre at point  $\hat{x}$ . The volume of this ellipsoid, equal to probability  $1 - \alpha$ , is controlled by the value of  $C_{\alpha}$ . This ellipsoid is known as the *confidence* (or error) *ellipsoid* of point  $P_j$ , or more accurately as a *point confidence ellipsoid*.

Let us enquire further into the nature of this ellipsoid. The system in which the coordinates are computed is not the most natural one for this ellipsoid because it does not coincide with its axes (cf. §3.1). A more convenient system in which to express the ellipsoid is that of the eigenvectors of  $C_{\hat{x}}$ , which are, of course, in the directions of the axes of the ellipsoid. Denoting by  $z_1, z_2, z_3$  the coordinates in this eigenvector system, in the ascending order of the eigenvalue magnitudes  $\sigma_1^{-2}, \sigma_2^{-2}, \sigma_3^{-2}$ , we get the equation of the point confidence ellipsoid in the following form:

$$[z_1, z_2, z_3] \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}^{-1} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = C_{\alpha}^2. \quad (16.22)$$

This equation, equivalent to (21), may also be written as

$$\frac{z_1^2}{C_{\alpha}^2 \sigma_1^2} + \frac{z_2^2}{C_{\alpha}^2 \sigma_2^2} + \frac{z_3^2}{C_{\alpha}^2 \sigma_3^2} = 1, \quad (16.23)$$

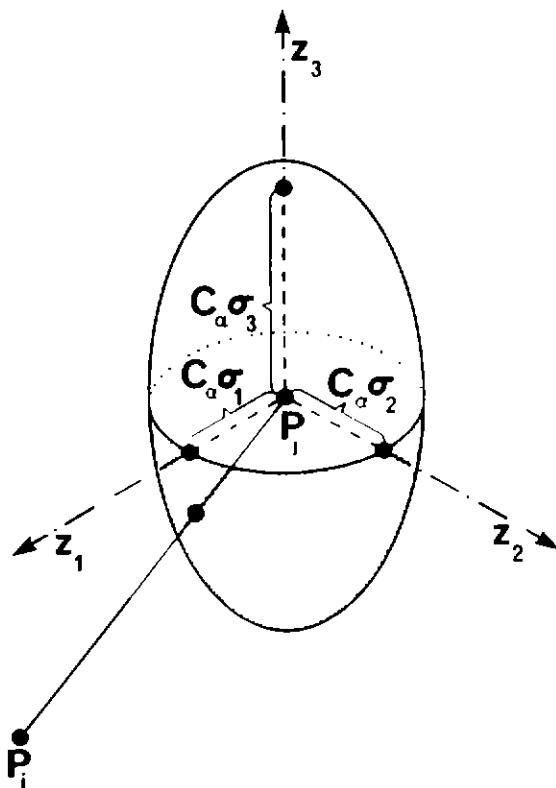


FIG. 16.7. Point confidence ellipsoid.

and is clearly the equation of a triaxial ellipsoid with minor semi-axis equal to  $C_\alpha \sigma_1$ , medium semi-axis equal to  $C_\alpha \sigma_2$ , and major semi-axis equal to  $C_\alpha \sigma_3$ , as shown in FIG. 7.

The relation between the expansion factor  $C_\alpha$  and the  $1 - \alpha$  probability that the correct position falls within the ellipsoid is (see (13.36))

$$C_\alpha = (\xi_{\chi^2_{\nu}, 1-\alpha})^{1/2}. \quad (16.24)$$

For  $C_\alpha = 1$ , which defines the *standard point confidence ellipsoid*, the corresponding probability  $1 - \alpha$  is 0.20. Conversely, if a probability of, say, 95% is desired, then the corresponding  $C_{0.05}$  is equal to 2.80.

All the preceding discussion applies to both the absolute and relative positioning. If the covariance matrix  $\mathbf{C}_x$  refers to point positioning, then some authors speak of an *absolute confidence ellipsoid*; if it refers to relative positioning, then it is called a *relative confidence ellipsoid*. In fact, there is nothing absolute about the confidence ellipsoid in the first case; the ellipsoid shows the degree of confidence, in the determined position, relative to the coordinate system implied by the technique used.

## 16.2. Relative horizontal positioning on reference ellipsoid

There are three kinds of observables used in relative horizontal positioning: astronomical azimuth  $A$ , horizontal angle  $\omega$  or direction  $d$ , and spatial distance  $\Delta r$ .

We will see that other observables enter into the model indirectly. If done properly, the computation of horizontal positions, using either a three-dimensional or two-dimensional approach, must yield identical results (up to rounding-off errors). When computing in three dimensions, as we did, for example, in the preceding section, the observations are not corrected, other than for instrumental effects and refraction, because the computations are carried out in the same space as the measurements. It is necessary to apply corrections to observations, however, before horizontal positions in two-dimensional spaces are computed; when computing on the reference ellipsoid, measurements made on the earth's surface must be reduced to this ellipsoid first. Reduction of observations to computational surfaces is an integral part of the direct problem of position determination. On the other hand, after the solution of the inverse problem on the ellipsoid, the derived quantities, i.e., the distance and azimuth, should be transformed back to the terrain using negative corrections. Accordingly, the topics of this section will be the reduction of observations onto the reference ellipsoid, mathematical models for the direct and inverse problems on the ellipsoid, the reduction of results of the inverse problem from the ellipsoid to the terrain, and, lastly, the assessment of the accuracy of relative positioning on the ellipsoid.

In the reduction of observations onto the ellipsoid, there are two groups of effects to be considered: geometrical effects, and the effects of the earth's gravity field. The geometrical effects arise from the peculiarities of the geometry of a biaxial ellipsoid. The gravity field must be considered because geodetic instruments used in the measurements are aligned to the gravity field—for instance, theodolites are aligned to the local plumb line—while computations are carried out in a geometrical space. The various reductions are functions of the position to be solved for. Thus the adoption of an iterative approach to corrections would appear necessary. However, because the corrections are very small, the first iteration is normally accurate enough.

Let us begin by examining the effect of the gravity field on the observed astronomical azimuth  $A$ . Recall that  $A$ , observed on the earth's surface, refers to the LA system (cf. FIG. 15.3) while a geodetic azimuth  $\alpha$  refers to the LG system (cf. FIG. 15.25). The difference between the two azimuths is given by the complete Laplace equation (15.90), which can be regarded as giving the *Laplace correction*  $\Delta\alpha = \alpha - A$ , also known as the correction for the deflection of the vertical, as

$$\Delta\alpha_{ij} = -\eta_i \tan \phi_i - (\xi_i \sin \alpha_{ij} - \eta_i \cos \alpha_{ij}) \cot Z_{ij} = +C_1 + C_2. \quad (16.25)$$

If the deflection is equal to zero, the correction is also equal to zero; for ideally horizontal lines ( $Z=90^\circ$ ), the second term is zero. A possible example of the Laplace correction is shown in FIG. 8. The corrected astronomical azimuth is simply the geodetic azimuth  $\alpha$  already defined in §15.4. The geodetic azimuth obtained in this way is often called the *Laplace azimuth*.

What is really needed for computations on the ellipsoid is an azimuth on the ellipsoid. Let us define it ( $\alpha'_{ij}$ ) first as the angle between the geodetic meridian plane and the plane given by the ellipsoidal normal at  $P_i$  and the projection  $P'_i$  and  $P_j$  onto

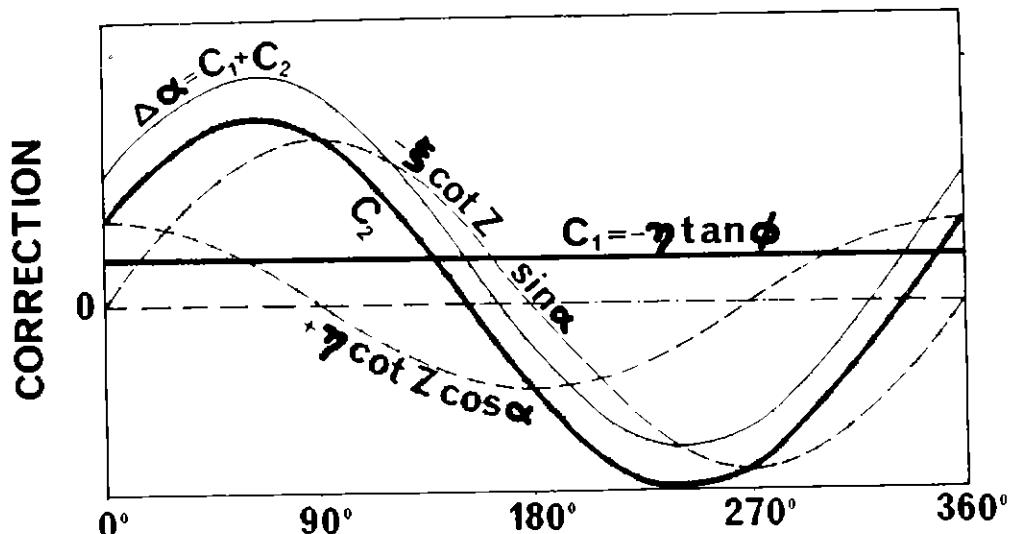


FIG. 16.8. An example of Laplace correction  $\Delta\alpha$ . (Assumed:  $\xi > 0$ ,  $\eta < 0$ ,  $z > \frac{1}{2}\pi$ ,  $\phi > 0$ .)

the ellipsoid (see FIG. 9). The latter plane intersects the ellipsoid in a normal section, and  $\alpha'_{ij}$  can be viewed as referring to it. The difference between  $\alpha'_{ij}$  and  $\alpha_{ij}$  is called the *skew-normal correction*; it clearly arises from the fact that the two ellipsoidal normals (at  $P_i$  and  $P_j$ ) are skewed and not coplanar. The correction is computed from [ZAKATOV, 1953]

$$\alpha'_{ij} - \alpha_{ij} = \frac{h_j}{2M_m} e^2 \sin 2\alpha_{ij} \cos^2 \phi_m, \quad (16.26)$$

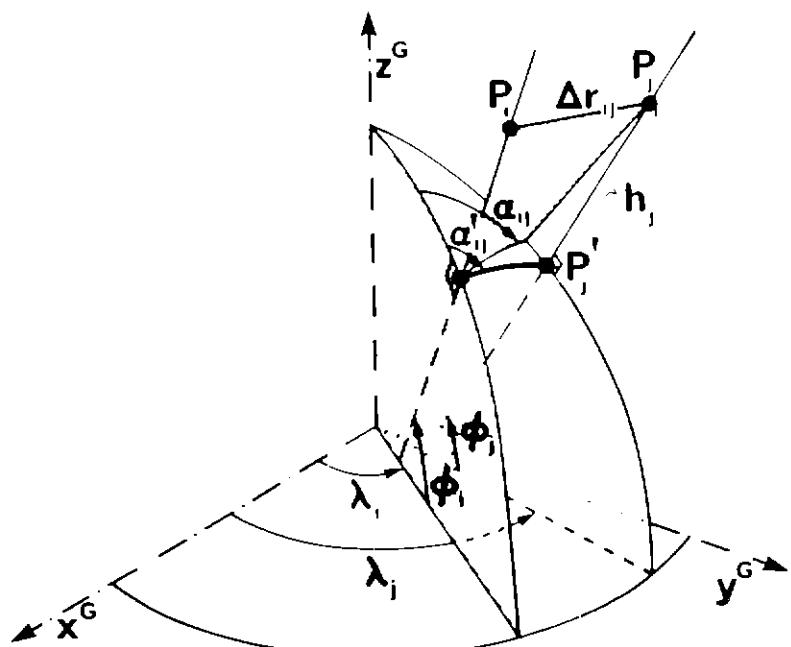


FIG. 16.9. Effect of skew normals.

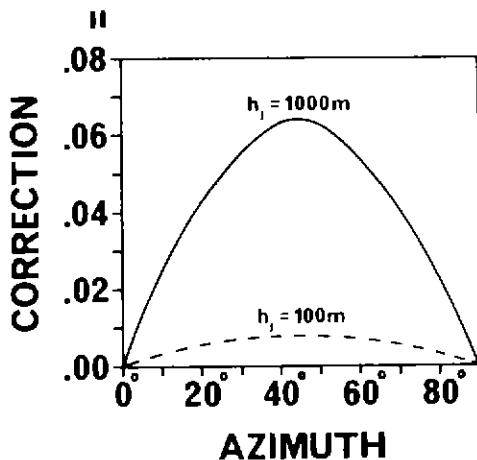


FIG. 16.10. An example of skew normal correction for  $\phi_i = 40^\circ$ ;  $\phi_j = 41^\circ$ .

where  $\phi_m = \frac{1}{2}(\phi_i + \phi_j)$ ,  $M_m = \frac{1}{2}(M_i + M_j)$ , and  $h_j$  is the height of  $P_j$  above the ellipsoid thereby explaining why the correction is sometimes called the *height of target correction*. Hence, one can write,

$$\alpha'_{ij} = \alpha_{ij} + C_3 = A_{ij} + C_1 + C_2 + C_3. \quad (16.27)$$

The dependence of  $C_3$  on  $\alpha_{ij}$  is shown in FIG. 10.

In a configuration of three points, for example, one can recognize six normal sections (cf. FIG. 11), thus the use of normal sections introduces an ambiguity in the definition of an ellipsoidal triangle. This problem is removed by the introduction of the geodesic curve (see §3.3), which is the curve that has the locally shortest length of all curves drawn on the ellipsoid between the given two end points. Its osculating plane contains the ellipsoidal normal at every point along its path. It is usually, but not always, bounded by the forward ( $P_i$  to  $P_j$ ) and reverse ( $P_j$  to  $P_i$ ) normal sections (cf. FIG. 11). The azimuth  $\alpha^E$  referring to the geodesic curve is called the *ellipsoidal azimuth*. The *normal section to geodesic correction* required to be added to the geodetic azimuth  $\alpha'$  of the forward normal section [BOMFORD, 1971] is given as

$$\alpha_{ij}^E - \alpha'_{ij} = -\frac{e^2 \Delta r_{ij}^2 \cos^2 \phi_m \sin 2\alpha_{ij}}{12 N_m^2}, \quad (16.28)$$

where all quantities have been previously defined except  $N_m = \frac{1}{2}(N_i + N_j)$ . The *complete azimuth correction* is

$$\alpha_{ij}^E - A_{ij} = C_1 + C_2 + C_3 + C_4. \quad (16.29)$$

The change in the value of  $C_4$  as a function of  $\Delta r$  and  $\alpha$  is given in FIG. 12.

Observed directions  $d$ , of which the azimuth is only a special kind, are reduced onto the ellipsoid by use of the same corrections. Thus, (29) applies to any observed direction and, as such, represents a general *horizontal direction correction*.

A horizontal angle  $\omega$  is reduced to the ellipsoid by applying (29) to the two directions of the arms of the angle. Clearly,  $C_1$  disappears (cf. (25)), and one is left

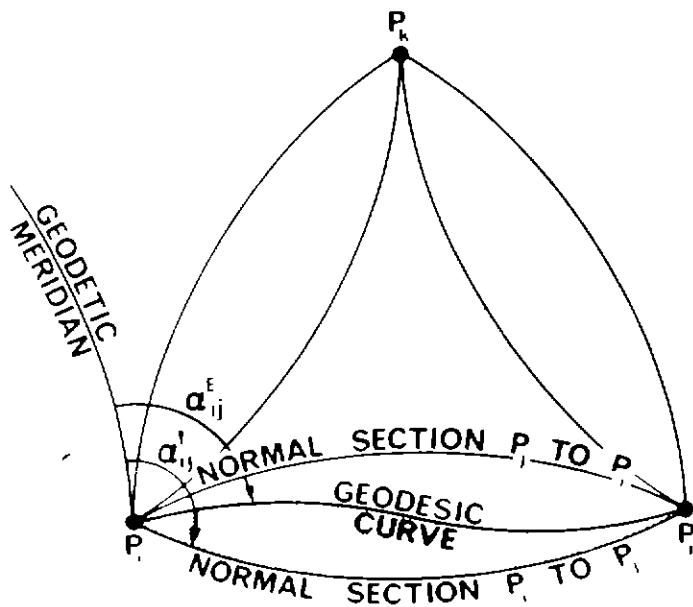


FIG. 16.11. Difference between normal sections and geodesic curve.

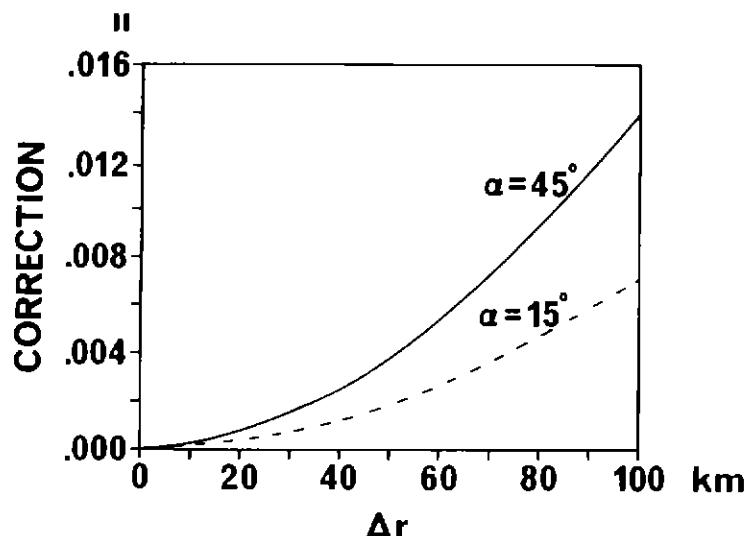


FIG. 16.12. Normal section to geodesic correction for mean latitude of \$45^\circ\$.

with the differences  $\Delta$  of the remaining three corrections. The result, the *horizontal angle correction*, can be written as

$$\omega^E - \omega = \Delta C_2 + \Delta C_3 + \Delta C_4. \quad (16.30)$$

Spatial distance  $\Delta r \equiv \rho$  is reduced to an ellipsoidal distance  $S^E$  (see FIG. 13) as follows [HEISKANEN AND MORITZ, 1967]:

$$S_{ij}^E = 2R \arcsin\left(\frac{l_{ij}^0}{2R_m}\right), \quad (16.31)$$

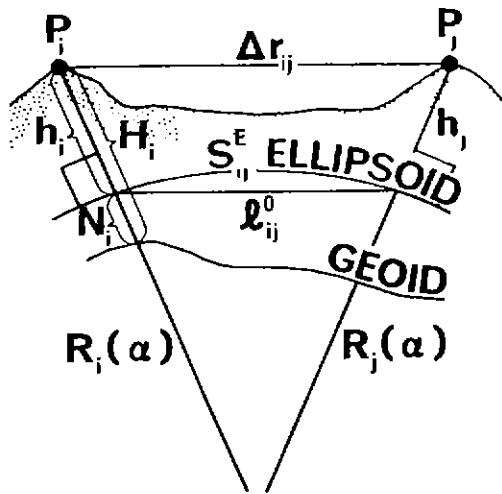


FIG. 16.13. Reduction of a spatial distance.

where

$$l_{ij}^0 = \sqrt{\frac{\Delta r_{ij}^2 - (h_j - h_i)^2}{(1 + h_i/R_m)(1 + h_j/R_m)}}, \quad (16.32)$$

$$R_m = \frac{1}{2}(R_i(\alpha) + R_j(\alpha)), \quad (16.33)$$

and the radius of curvature in the azimuth  $\alpha = \alpha_{ij}$ , is given by (3.89), written here as

$$R_i(\alpha) = \frac{M_i N_i}{M_i \sin^2 \alpha + N_i \cos^2 \alpha}. \quad (16.34)$$

The relative *distance correction*  $(S^E - \Delta r)/\Delta r$  is approximately  $10^{-6}$  for every 6.4 m of height. Shown in FIG. 14 is the error in distances if in the reduction the orthometric instead of the geodetic height is used [VANÍČEK AND MERRY, 1973]. Clearly, for a first-order network, the geoidal heights have to be taken into consideration. On the other hand, the difference in the lengths of the normal section and geodesic curve does not have to be taken into account; it reaches at most  $0.74 \times 10^{-5}$  m for a line of 600 km long [ZAKATOV, 1953].

In many cases, the magnitudes of the above corrections are small, and there is a temptation not to apply them. One has to realize, however, that since the accumulation of systematic errors has a more harmful effect than that of random errors, even the smallest neglected corrections may lead to a significant accumulation of distortions if points are strung together, as will be shown in §18.3. It is left to the geodesist to judge if these corrections are significant for any given project.

Once the observations are reduced to the ellipsoid, it is possible to formulate the *direct problem on the ellipsoid*: Given  $(\phi_i, \lambda_i)$  of point  $P_i$ , the ellipsoidal azimuth  $(\alpha_{ij}^E)$  and ellipsoidal distance  $(S_{ij}^E)$  to point  $P_j$ , compute  $(\phi_j, \lambda_j)$  of  $P_j$ . There exists a multitude of solutions. They can be broken down into three families according to which curve they use: geodesic, normal section, or normal section on the local

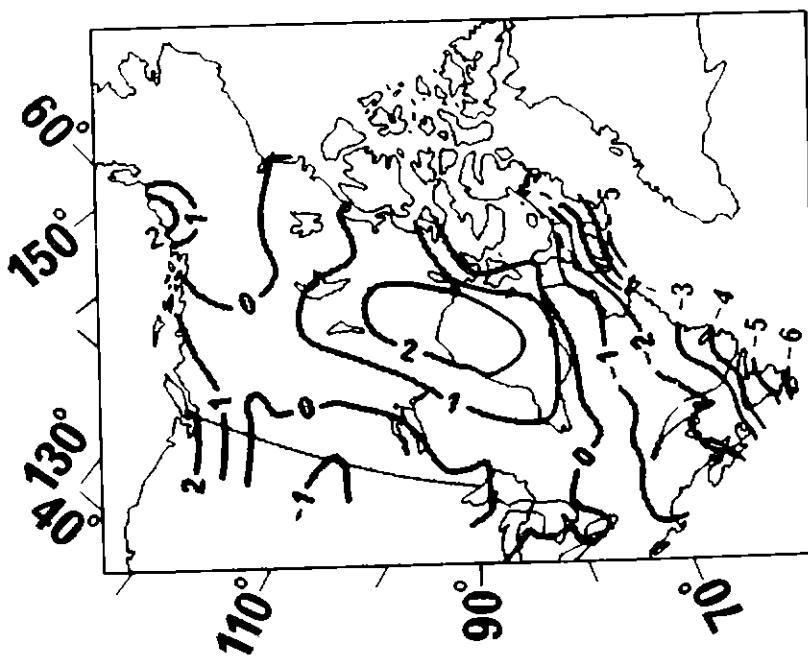


FIG. 16.14. Effect of neglecting the geoidal height  $N$  in distance reduction in Canada. Contours in units of  $10^{-6}$ .

spherical approximation of the ellipsoid. Before presenting any formulae, let us give an overview of these three families of solutions.

The most accurate solutions are those known as *long line formulae* based on the use of geodesic curves; they require the evaluation of an elliptic integral (§3.2). Of these, we should mention at least Bessel's [JORDAN AND EGGERT, 1962], RAINSFORD's [1955], and SODANO's [1965] methods. The accuracy of any of these three methods is limited only by the number of terms one wishes to take in the series involved. The only difference between Bessel's and Rainsford's approaches is that the latter uses  $f$  instead of  $e^2$ . This results in a more rapid convergence of the iterative procedure that, one way or another, must be used. The main advantage of the third method is that both the direct and inverse problems are solved by Sodano in a non-iterative fashion; the formulae have been iterated internally so they can be used directly. Other closed formulae may be found in THOMAS [1972].

The other two families contain approximate formulae and, as such, should not be used for either longer lines or all locations. The formulae that use the normal section on the ellipsoid can be found in, e.g., ROBBINS [1962], and those that utilize the local spherical approximation, known as the *short line formulae*, are given in, e.g., BOMFORD [1971], and elsewhere. Because of their availability and because of the tediousness of their derivations, we give here only the most widely used one, *Puissant's formula*. This solution is good to a relative accuracy of  $10^{-6}$  for a line 100 km long, but the accuracy deteriorates rapidly to more than  $40 \times 10^{-6}$  for a line 250 km long for  $\phi \geq 60^\circ$ .

The Puissant solution is sought in the form of latitude difference first; namely, omitting superscripts E by both  $S$  and  $\alpha$ :

## RELATIVE POSITIONING

$$\Delta\phi^{(k+1)} = \left( \frac{S_{ij} \cos \alpha_{ij}}{M_i} - \frac{S_{ij}^2 \tan \phi_i \sin^2 \alpha_{ij}}{2M_i N_i} \right. \\ \left. - \frac{S_{ij}^3 \cos \alpha_{ij} \sin^2 \alpha_{ij} (1 + 3 \tan^2 \phi_i)}{6M_i N_i^2} \right) \left( 1 - \frac{3e^2 \sin 2\phi_i}{4(1 - e^2 \sin^2 \phi_i)} \Delta\phi^{(k)} \right). \quad (16.35)$$

Since, in the above equation,  $\Delta\phi = \Delta\phi_{ij}$ , is also present in the corrective term on the right-hand side, the solution  $\Delta\phi$  can be obtained only in an iterative manner. Iterations are usually continued until  $|\Delta\phi^{(k+1)} - \Delta\phi^{(k)}|$  falls below a certain limit, say,  $10^{-9}$  rad, that corresponds to an uncertainty of about 6 mm in the horizontal position. The resultant latitude is then given as  $\phi_j = \phi_i + \Delta\phi^{(k+1)}$ , the longitude as  $\lambda_j = \lambda_i + \Delta\lambda$ , where  $\Delta\lambda$  is evaluated directly from

$$\Delta\lambda = \frac{S_{ij}}{N_j} \frac{\sin \alpha_{ij}}{\cos \phi_j} \left[ 1 - \frac{S_{ij}^2}{6N_j^2} \left( 1 - \frac{\sin^2 \alpha_{ij}}{\cos^2 \phi_j} \right) \right]. \quad (16.36)$$

Computation of the ellipsoidal azimuth  $\alpha_{ji}^E$  of the line reckoned from  $P_j$  to  $P_i$ , called the *inverse azimuth*, is usually regarded as part of the direct problem and is carried out by means of the following formula:

$$\alpha_{ji}^E - \alpha_{ij}^E - \pi = \Delta\lambda \frac{\sin \phi_m}{\cos \frac{1}{2}\Delta\phi} + \frac{\Delta\lambda^3}{12} \left[ \frac{\sin \phi_m}{\cos \frac{1}{2}\Delta\phi} - \frac{\sin^3 \phi_m}{\cos^3 \frac{1}{2}\Delta\phi} \right]. \quad (16.37)$$

Deletion of higher than first order terms results in *Gauss's mid-latitude formula* which is valid only for lines up to 40 km long [ALLAN ET AL., 1968]. The approach based on observing two astronomical azimuths  $A_{1j}$  and  $A_{2j}$ , or two spatial distances  $\Delta r_{1j}$  and  $\Delta r_{2j}$  (instead of one azimuth and one direction) from two known points  $P_1$  and  $P_2$  to the unknown point  $P_j$  can also be used and is discussed in detail in §18.4.

Puissant's solution to the *inverse problem on the ellipsoid* reads as follows:

$$\alpha_{ij}^E = \arctan \left[ \frac{N_j \Delta\lambda}{M_i \Delta\phi} \cos \phi_j \left( 1 - \frac{3e^2 \sin 2\phi_i}{4(1 - e^2 \sin^2 \phi_i)} \right) \right], \quad (16.38)$$

and

$$S_{ij}^E = \frac{\Delta\phi}{\cos \alpha_{ij}} \frac{M_i}{1 - \frac{3e^2 \sin 2\phi_i \Delta\phi}{4(1 - e^2 \sin^2 \phi_i)}}. \quad (16.39)$$

The distance and azimuth can then be transformed to the earth's surface by applying

the reduction formulae ((29) and (31)) in the reversed sense.

The accuracy of relative position on the ellipsoid is assessed in a manner parallel to that of the three-dimensional case: the equations found at the end of §16.1 apply here merely by dropping one dimension. This time we have only a pair of coordinates  $\hat{x} = (\hat{\phi}, \hat{\lambda})^T$  along with the corresponding covariance matrix  $C_{\hat{x}}$  of dim(2, 2), which can be interpreted as a *point confidence ellipse*. The expansion factor  $C_\alpha$  for the  $1 - \alpha$  point confidence ellipse is defined as (cf. §13.5)

$$C_\alpha = (\xi_{\chi^2_{1-\alpha}})^{1/2}. \quad (16.40)$$

When  $C_\alpha = 1$ , the corresponding probability is 0.39, which defines the *standard point confidence ellipse*. Conversely, for a desired probability of, say, 95%, the corresponding  $C_{0.05}$  is equal to 2.45, a little less than the figure for the three-dimensional case (§16.1). As in the three-dimensional case, we speak of *relative confidence ellipse* and *absolute confidence ellipse*, whose meanings should be clear by now.

### 16.3. Relative horizontal positioning on conformal map

In §15.4, the concept of mapping was introduced within the context of the transformation of positions; we now take a more detailed look at it here. A mapping of a closed domain  $\mathfrak{D}_1$  on a surface  $S_1$  onto a closed domain  $\mathfrak{D}_2$  on a second surface  $S_2$  is given by (15.100) or (15.101), which we shall write here simply as

$$\mathbf{x} = \mathbf{x}(\mathbf{u}). \quad (16.41)$$

The coordinates  $(u, v)$  are defined on  $S_1$  and  $(x, y)$ —where the superscript M has been dropped—are on  $S_2$ . By means of (41), points  $P(u, v) \in \mathfrak{D}_1$  are related to points  $P(x, y) \in \mathfrak{D}_2$ , and the region  $\mathfrak{D}_1 \subset S_1$  is thus said to be mapped onto the region  $\mathfrak{D}_2 \subset S_2$ . To be useful in geodesy, the mapping equations (41) must fulfil the following conditions: they must be unique, finite, twice differentiable at all points in  $\mathfrak{D}_1$ , and the Jacobian  $|\partial \mathbf{x} / \partial \mathbf{u}|$  (see §3.1) must differ from zero in  $\mathfrak{D}_1$ . A transformation with these properties (except for the double differentiability) is said to be a *diffeomorphic transformation*.

Let us now apply the diffeomorphic transformations to the mapping of the ellipsoid ( $S_1$ ) onto a map ( $S_2$ ). In FIG. 15, a mapping of an ellipsoidal triangle  $P_i P_j P_k$  is shown: note how the shape of the triangle and geodesics between the pairs of points get distorted through the mapping. The mapped geodesics are called *projected geodesics*. The length  $S^M$  of the projected geodesic is generally not the same as the length  $S^E$  of the original geodesic on the ellipsoid. The projected geodesic is usually not a straight line, i.e., it is not a geodesic curve on the mapping plane; also, a pair of projected geodesics do not ordinarily intersect under the same angle as those on the ellipsoid. This is why the *ellipsoidal angle*  $\omega^E$ , defined as the angle between the tangents to the two geodesics on the ellipsoid, is not equal to the

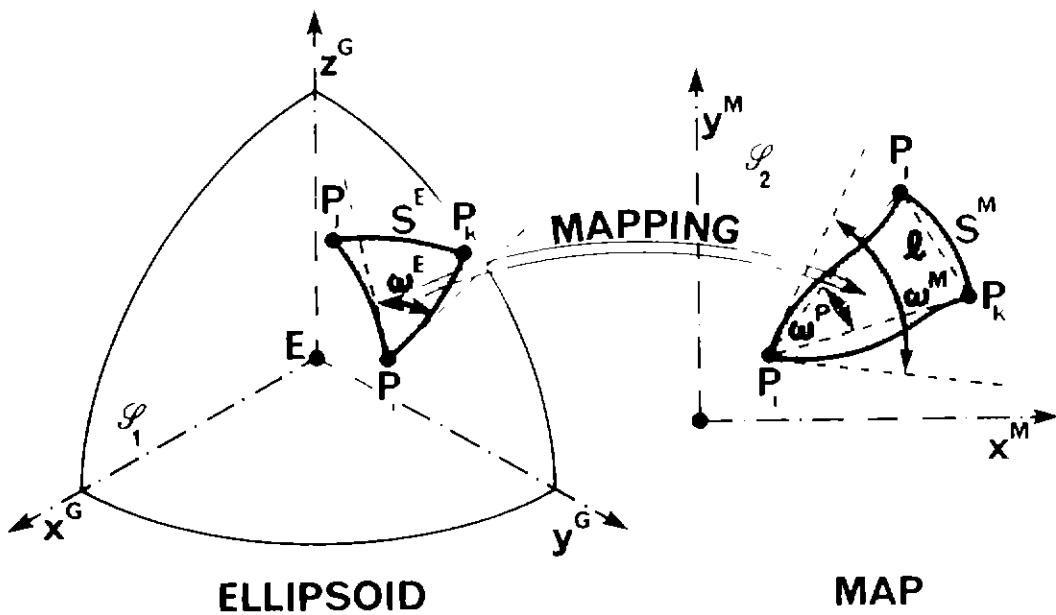


FIG. 16.15. Mapping of ellipsoid onto a plane.

corresponding *projected angle*  $\omega^M$ , defined as the angle between the tangents to two projected geodesics (see FIG. 15). The third angle needed is that between the two chords, called the *plane angle*  $\omega^P$ ; it is this angle that is used in the computations on the map, together with the *chord length*  $l$ . Thus, referring to reductions of observables from the earth's surface onto the map, we want to end up with angles  $\omega^P$  and distances  $l$ .

Let us begin with an examination of the length distortion. The amount of length distortion in a specific direction at a point on the mapping plane is described by the *point scale factor*  $k$ . It is defined by the following ratio:

$$k = \frac{dS^M}{dS^E}. \quad (16.42)$$

Here,  $dS^E$  is a length differential on the ellipsoid that can be defined as

$$dS^E = \psi_E(u, v) \sqrt{du^2 + dv^2}, \quad (16.43)$$

where  $\psi_E$  is a scale function of  $u, v$  that generally changes with position; and  $u, v$  are some parameters (see §3.3) on the ellipsoid. A length differential  $dS^M$  on the map is defined as

$$dS^M = \psi_M(x, y) \sqrt{dx^2 + dy^2}, \quad (16.44)$$

where  $\psi_M$  is a scale function of  $x, y$ . The role of the scale functions will be clarified later. By expressing the differentials  $dx, dy$  in (44) in terms of  $du, dv$  from (41) and then substituting (43) and (44) into (42), we get

$$k^2 = \frac{\psi_M^2(x, y)}{\psi_E^2(u, v)} \frac{(e du^2 + 2f du dv + g dv^2)}{du^2 + dv^2}, \quad (16.45)$$

where  $e$ ,  $f$ , and  $g$  are the Gaussian fundamental quantities defined as (cf. §3.3)

$$e = \left( \frac{\partial x}{\partial u} \right)^2 + \left( \frac{\partial y}{\partial u} \right)^2, \quad (16.46)$$

$$f = \frac{\partial x}{\partial u} \frac{\partial x}{\partial v} + \frac{\partial y}{\partial u} \frac{\partial y}{\partial v}, \quad (16.47)$$

$$g = \left( \frac{\partial x}{\partial v} \right)^2 + \left( \frac{\partial y}{\partial v} \right)^2. \quad (16.48)$$

In a more compact form, we have

$$k^2 = \frac{\psi_M^2}{\psi_E^2} \frac{[du, dv] \mathbf{G} \begin{bmatrix} du \\ dv \end{bmatrix}}{[du, dv] \begin{bmatrix} du \\ dv \end{bmatrix}}, \quad (16.49)$$

where  $\mathbf{G}$  is given by (3.87). Clearly, (49) is an equation of an ellipse with the direction of maximum point scale factor being associated with the major axis, and the minimum point scale factor being associated with the minor axis. The size, shape, and orientation of this ellipse is obtained in the usual manner by transforming (49) into its eigenvector coordinate system (§3.1). The geometrical interpretation of (49) is a more general form of Tissot's indicatrix (see §3.3) which includes the ratio of the scale functions. For the mappings used in geodesy, the indicatrix is elliptical in shape.

Of special interest may be the value of  $k$  in a particular direction. This direction is usually reckoned clockwise from the direction  $y^T$  of maximum scale—see FIG. 16—either on the ellipsoid ( $t$ ) or on the map ( $t'$ ). While the Tissot indicatrix is the locus of  $k(t')$ , the locus of  $k(t)$  is the *pedal curve* to the Tissot indicatrix. Mathematically, the pedal curve is given by

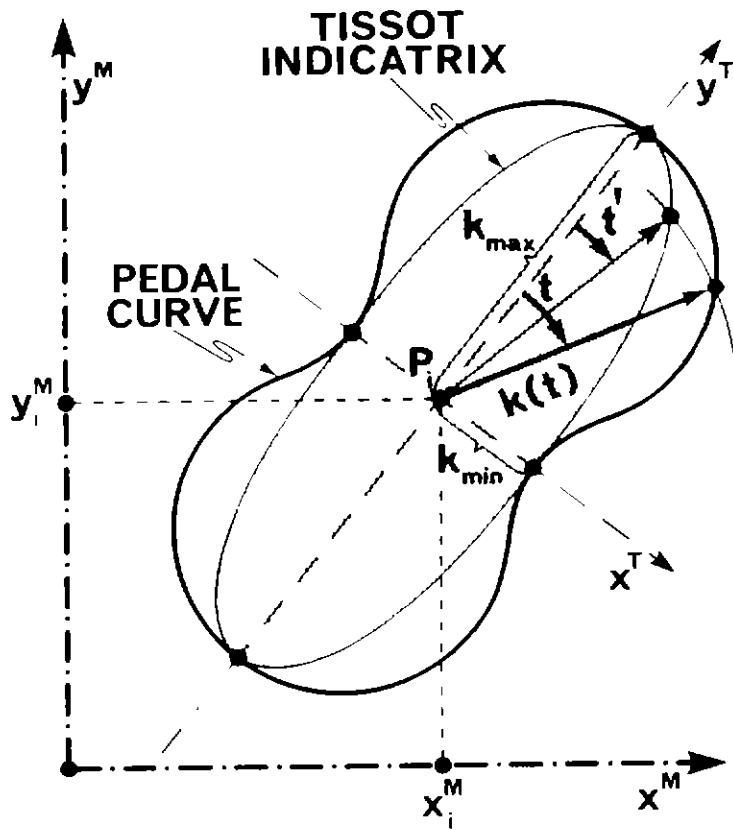
$$k^2(t) = \frac{\psi_M^2}{\psi_E^2} \mathbf{q}^T(t) \mathbf{G} \mathbf{q}(t), \quad (16.50)$$

where

$$\mathbf{q}(t) = [\cos t, \sin t]^T. \quad (16.51)$$

We are now finally in a position to define a conformal mapping and to show the special importance of conformality. A *conformal mapping* is a mapping for which the point scale  $k$  is a function only of position and not azimuth, i.e., for which  $k$  is isotropic. Clearly, under these conditions, the Tissot indicatrix as well as its pedal curve are circular. The reader can satisfy himself (see (45)) that the definition leads to the following (necessary and sufficient) *conditions for conformality*:

$$f = 0, \quad e = g. \quad (16.52)$$

FIG. 16.16. Tissot's indicatrix and pedal curve at point  $P_i$ .

An alternative and equivalent statement of the conditions for conformality are the *Cauchy-Riemann equations* that can be deduced from (46) to (48) [HOTINE, 1946; 1947] for the  $(u, v)$  and  $(x, y)$  systems being of the same handedness,

$$\frac{\partial x}{\partial u} = \frac{\partial y}{\partial v} \quad \text{and} \quad \frac{\partial x}{\partial v} = -\frac{\partial y}{\partial u}, \quad (16.53)$$

or for opposite handedness,

$$\frac{\partial x}{\partial v} = \frac{\partial y}{\partial u} \quad \text{and} \quad \frac{\partial x}{\partial u} = -\frac{\partial y}{\partial v}. \quad (16.54)$$

The projected angle  $\omega^M$  at every point of a conformal map is equal to its corresponding angle  $\omega^E$  on the ellipsoid. We then say that conformal projections preserve angles; as a matter of fact they preserve the form of infinitesimally small configurations—hence the name. Since, as will be shown later, it is a relatively simple matter to obtain the plane angle  $\omega^P$  from the projected angle  $\omega^M$ , it is possible to make an easy transition from a horizontal angle  $\omega$  measured on the earth's surface to the plane angle  $\omega^P$  on a conformal map through  $\omega^E = \omega^M$ , and vice versa. This is clearly a big advantage that is not shared by non-conformal projections; thus the use of conformal projections in geodesy is strictly a matter of computational convenience.

ence. It should be noted that the selection of an ideal map projection for cartographic purposes is another matter altogether, and the appropriate literature should be consulted, e.g., MALING [1973].

To be able to use the formulae derived above as directly as possible, it is expedient to introduce a special parameter  $q$ , called the *isometric latitude*, on the ellipsoid. Let us first consider the length differential  $dS^E$  on the ellipsoid. According to FIG. 17, we get

$$(dS^E)^2 = (Md\phi)^2 + (N \cos \phi d\lambda)^2 = N^2 \cos^2 \phi \left( \frac{M^2}{N^2} \frac{d\phi^2}{\cos^2 \phi} + d\lambda^2 \right). \quad (16.55)$$

By defining the differential of the new parameter as

$$dq = \frac{M}{N} \frac{d\phi}{\cos \phi}, \quad (16.56)$$

(55) may be rewritten as

$$(dS^E)^2 = N^2 \cos^2 \phi (dq^2 + d\lambda^2). \quad (16.57)$$

It can be seen that this equation conforms to (43), if  $u = q$ ,  $v = \lambda$ . The isometric latitude  $q$  is clearly defined only in terms of the geodetic latitude  $\phi$  and is obtained by integrating (56). It represents simply an alternative parameter on the ellipsoid, replacing the geodetic latitude  $\phi$ , that allows us to directly use (45). We obtain [HOTINE, 1946]

$$\begin{aligned} q &= \int_0^\phi \frac{M}{N \cos \xi} d\xi = \ln \tan\left(\frac{1}{4}\pi + \frac{1}{2}\phi\right) \left( \frac{1 - e \sin \phi}{1 + e \sin \phi} \right)^{e/2} \\ &= \operatorname{arctanh}(\sin \phi) - e \operatorname{arctanh}(e \sin \phi), \end{aligned} \quad (16.58)$$

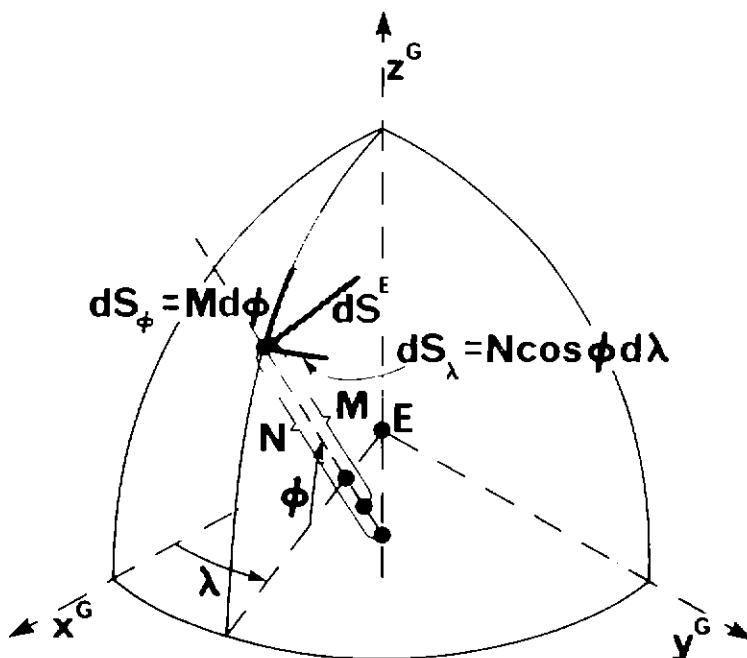


FIG. 16.17. Length differential on reference ellipsoid.

TABLE 16.1  
Variations of isometric latitude

Latitude $\phi$	Isometric latitude $q$
$< 11^\circ$	$< \phi$
$< \phi \leq 90^\circ$	$> \phi$
$\rightarrow 90^\circ$	$\rightarrow \infty$

where, here,  $e$  is the eccentricity of the reference ellipsoid. Direct computations show the values of  $q$  given in TABLE 1.

If the parameters  $q, \lambda$  are viewed as spanning a plane, called the *isometric plane*, then the ellipsoidal parallels are mapped as straight lines with varying spacing. Meridians are mapped perpendicular to the parallels and equally spaced. The isometric plane is sometimes called Mercator's projection.

A conformal mapping of an ellipsoid must clearly satisfy the following equations (cf. (53) and (54)):

$$\frac{\partial x}{\partial q} = \frac{\partial y}{\partial \lambda} \quad \text{and} \quad \frac{\partial y}{\partial q} = -\frac{\partial x}{\partial \lambda}, \quad (16.59)$$

for  $(q, \lambda)$  and  $(x, y)$  having the same handedness, or

$$\frac{\partial x}{\partial q} = -\frac{\partial y}{\partial \lambda} \quad \text{and} \quad \frac{\partial y}{\partial q} = \frac{\partial x}{\partial \lambda}, \quad (16.60)$$

for opposite handedness. For any conformal mapping we then get (see (49))

$$k^2(q, \lambda) = \frac{1}{N^2 \cos^2 \phi} \frac{e d q^2 + e d \lambda^2}{d q^2 + d \lambda^2} = \frac{1}{N^2 \cos^2 \phi} \frac{g d q^2 + g d \lambda^2}{d q^2 + d \lambda^2}, \quad (16.61)$$

where  $e$  and  $g$  are the fundamental quantities. These equations can be rewritten as

$$k(q, \lambda) = \frac{\sqrt{\left(\frac{\partial x}{\partial \lambda}\right)^2 + \left(\frac{\partial y}{\partial \lambda}\right)^2}}{N \cos \phi} = \frac{\sqrt{\left(\frac{\partial x}{\partial q}\right)^2 + \left(\frac{\partial y}{\partial q}\right)^2}}{N \cos \phi}. \quad (16.62)$$

Another quantity necessary for the characterization of conformal mappings is the convergence of the projected geodetic meridians on the mapping plane. The *meridian convergence*  $\gamma$  is defined as the angle between the tangent to the projected meridian and the  $y$ -axis of the projection (FIG. 18). To obtain a formula for  $\gamma$ , we first write the general expression for the projected meridian as

$$y = y(x). \quad (16.63)$$

Its slope is

$$\cot \gamma = -\frac{dy}{dx}, \quad (16.64)$$

and, expressing  $dy, dx$  as total differentials of the mapping equations, we have

$$\cot \gamma = - \frac{\frac{\partial y}{\partial q} dq + \frac{\partial y}{\partial \lambda} d\lambda}{\frac{\partial x}{\partial q} dq + \frac{\partial x}{\partial \lambda} d\lambda}. \quad (16.65)$$

Realizing now that along the geodetic meridian  $\lambda = \text{const.}$ , and thus  $d\lambda = 0$ , we finally obtain

$$\cot \gamma = - \frac{\partial y}{\partial q} / \frac{\partial x}{\partial q} = \frac{\partial x}{\partial \lambda} / \frac{\partial y}{\partial \lambda}, \quad (16.66)$$

where the second equation is obtained by employing the Cauchy–Riemann equations. For example, the meridian convergence of Mercator's projection is equal to zero.

Let us now describe in detail the geometry of the curves projected from the ellipsoid onto the conformal plane. Aside from the geodetic meridian discussed above, FIG. 18 also shows the projected geodesic (joining the projected positions of points  $P_i, P_j$ ) and their relationship to the mapping coordinate system. The *projected azimuth*  $\alpha^M$  is the angle between the tangent to the projected meridian and the

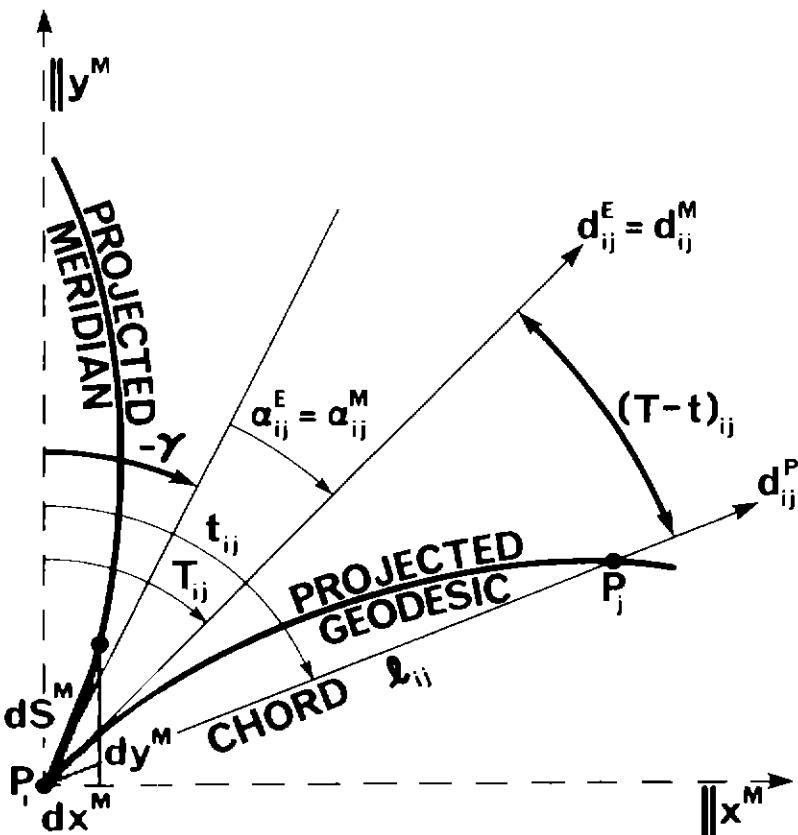


FIG. 16.18. Meridian convergence and  $T - t$  correction.

tangent to the projected geodesic. Of course, it identically equals the ellipsoidal azimuth  $\alpha^E$ , since the projection is conformal. The *grid azimuth* ( $T$ ) of the projected geodesic is the angle between *grid north*, i.e., the  $y^M$ -axis, and the tangent to the projected geodesic. The *grid azimuth of the chord* ( $t$ ) is the angle between grid north and the chord of  $P_i, P_j$ .

If the computations on the conformal map are to be simple, they must utilize the chords and all the quantities related to them. The quantities related to the chord are obtained as follows:

(a) An ellipsoidal azimuth is reduced to the grid azimuth by first subtracting the meridian convergence (66):

$$T_{ij} = \alpha_{ij}^E - \gamma_i. \quad (16.67)$$

The grid azimuth is further reduced by the *arc-to-chord correction*, often called just a  *$T - t$  correction*, to give the grid azimuth of the chord:

$$t_{ij} = T_{ij} - (T - t)_{ij}. \quad (16.68)$$

The development of specific expressions for the  $T - t$  correction is complex. It involves working with the parametric equations of the projected geodesic and evaluating the curvature of this curve. The interested reader is referred to THOMAS [1952].

(b) An ellipsoidal direction  $d^E$  is reduced to the plane direction  $d^P$  by the same correction (see FIG. 18):

$$d_{ij}^P = d_{ij}^E - (T - t)_{ij}. \quad (16.69)$$

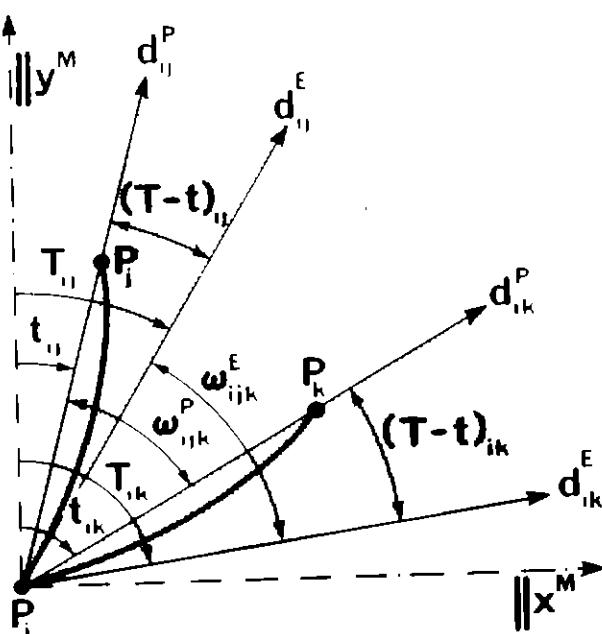


FIG. 16.19. Reduction of angles and directions.

(c) An ellipsoidal angle  $\omega^E$  is reduced by (see FIG. 19)

$$\omega_{ijk}^P = \omega_{ijk}^E + (T - t)_{ij} - (T - t)_{ik}. \quad (16.70)$$

(d) A length of the geodesic on the ellipsoid,  $S^E$ , is transformed to the length of the projected geodesic by the *line scale*  $\bar{k}$ , and one gets

$$l_{ij} = S_{ij}^M = \bar{k} S_{ij}^E, \quad (16.71)$$

where the line scale is the average point scale  $k$  over the line. For short and medium length lines, the difference in length between the projected geodesic and the chord is so minute it does not merit explicit treatment.

Naturally, the expressions for  $\gamma$ ,  $T - t$ , and  $k$  vary from one conformal mapping to another. To treat these individual mappings here is considered beyond the scope of this book. The reader is referred to LEE [1976].

Once all the needed observables have been reduced into the conformal mapping plane, relative positions can be computed. The *direct problem on the mapping plane* is expressed by the following straightforward equations:

$$x_j^M = x_i^M + l_{ij} \sin t_{ij}, \quad y_j^M = y_i^M + l_{ij} \cos t_{ij}. \quad (16.72)$$

The solution is obtained through iterations as both the line scale  $\bar{k}$  and the  $(T - t)$  correction needed to get  $l_{ij}$  and  $t_{ij}$  are also functions of  $x_j^M$  and  $y_j^M$ . The first approximation  $(x_j^M)^{(1)}, (y_j^M)^{(1)}$  is computed from (72) using  $S_{ij}$  and  $T_{ij}$  instead of  $l_{ij}$  and  $t_{ij}$ . The accuracy of the relative positions  $x_j^M, y_j^M$  is assessed the same way as in §16.2.

The *inverse problem on the mapping plane* is also given by straightforward expressions:

$$l_{ij} = \sqrt{(x_j^M - x_i^M)^2 + (y_j^M - y_i^M)^2}, \quad (16.73)$$

and

$$t_{ij} = 2 \arctan \frac{x_j^M - x_i^M}{y_j^M - y_i^M + \sqrt{(x_j^M - x_i^M)^2 + (y_j^M - y_i^M)^2}}. \quad (16.74)$$

These quantities refer, of course, to the mapping plane and can then be reduced to the ellipsoid (i.e., to  $S_{ij}^E$  and  $\alpha_{ij}^E$ ) through formulae inverse to (71), (68), and (67). The reductions from the ellipsoid up to the terrain (to get  $\Delta r_{ij}$  and  $A_{ij}$ ) were discussed in §16.2. Thus a completely closed system of reductions, i.e., terrain–ellipsoid–mapping plane–ellipsoid–terrain, is obtained.

Occasionally, it may be expedient to also use the transformation,  $q \rightarrow \phi$ , from the isometric plane to the ellipsoid. This indicates a transformation inverse to (58);

however, since (58) cannot be inverted, i.e.,  $\phi$  cannot be expressed explicitly as a function of  $q$ , an iterative technique has to be used. For instance, the Newton-Raphson iteration method (e.g., CONTE AND DE BOOR [1972]) yields

$$\phi^{(k)} = \phi^{(k-1)} - \frac{f(\phi^{(k-1)})}{f'(\phi^{(k-1)})} = \phi^{(k-1)} - (\Delta\phi)^{(k-1)}. \quad (16.75)$$

From (58)

$$f(\phi) = \frac{1}{2} [\ln(1 + \sin \phi) - \ln(1 - \sin \phi) + e \ln(1 - e \sin \phi) - e \ln(1 + e \sin \phi)] - q, \quad (16.76)$$

where  $e$  is the eccentricity of the reference ellipsoid, and

$$f'(\phi) = \frac{d f(\phi)}{d \phi} = \frac{(1 - e^2)}{(1 - e^2 \sin^2 \phi) \cos \phi}. \quad (16.77)$$

The iterative process begins by approximating (58) with

$$q = \ln \tan\left(\frac{1}{4}\pi + \frac{1}{2}\phi^{(0)}\right), \quad (16.78)$$

and thus obtaining

$$\phi^{(0)} = 2 \arctan \exp(q) - \frac{1}{2}\pi. \quad (16.79)$$

The iterations are continued until  $|\Delta\phi|^{(k-1)} < \epsilon$ , where  $\epsilon$  is some a priori chosen value. For  $\epsilon = 10^{-12}$ , corresponding to an accuracy in position better than 0.1 mm, convergence is achieved within about three iterations for any  $|\phi| < 89^\circ$ .

#### 16.4. Relative vertical positioning

Relative vertical positioning is concerned with the determination of the vertical position (i.e., the height) of one point with respect to another, i.e., with the determination of height difference. Only some of the known techniques will be discussed here; the rest of the techniques appear to be more appropriately treated in §19.4, as they are better suited for determining continuous height profiles.

*Trigonometrical height difference determination* requires zenith distances  $Z_{ij}$  and  $Z_{ji}$  observed with a geodetic theodolite at points  $P_i$  and  $P_j$ , as well as the knowledge of the deflections of the vertical  $\theta_i$  and  $\theta_j$  projected onto the plane given by  $C$ ,  $P_i$ , and  $P_j$ , where  $C$  is the mid-point of the ellipsoidal normals which are  $R_i$  and  $R_j$  long (see FIG. 20). The projection  $\epsilon_{ij}$  of the deflection of the vertical  $\theta_i$  onto the plane  $CP_iP_j$  with azimuth  $\alpha_{ij}$ , is evaluated from the deflection components as

$$\boxed{\epsilon_{ij} = \xi_i \cos \alpha_{ij} + \eta_i \sin \alpha_{ij},} \quad (16.80)$$

as the reader can see from FIG. 21 (cf. (15.89)). The geodetic zenith distance  $Z'_{ij}$

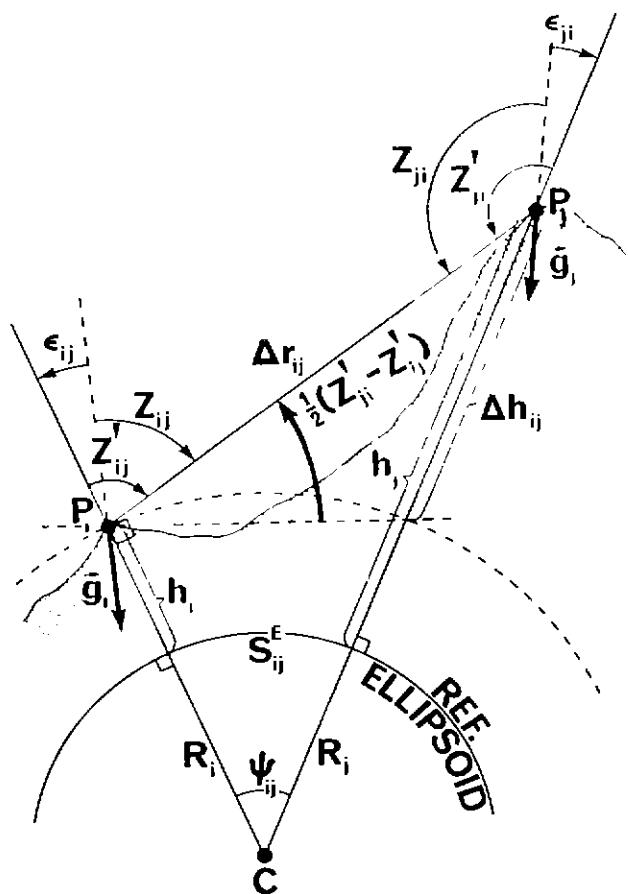


FIG. 16.20. Trigonometrical height difference.

referred to the ellipsoidal normal is thus determined from

$$Z'_{ij} = Z_{ij} + \epsilon_{ij} = Z_{ij} + \xi_i \cos \alpha_{ij} + \eta_i \sin \alpha_{ij}, \quad (16.81)$$

and similarly for  $Z'_{ji}$ . Note that in FIG. 20, the signs of  $\epsilon_{ij}$  and  $\epsilon_{ji}$  are clearly positive.

From the geodetic zenith distances, the *geodetic height* difference (from the ellipsoid)  $\Delta h_{ij} = h_j - h_i$  can be determined. For short lines of up to about 10 km, it is sufficiently accurate [HEISKANEN AND MORITZ, 1967] to regard the geodesic  $S^E$  between points  $P_i$  and  $P_j$  as a spherical arc with radius  $R_m = \frac{1}{2}(R_i + R_j)$ , where  $R_i$  and  $R_j$  are the radii of the ellipsoid at  $P_i$ ,  $P_j$  in the azimuth  $\alpha_{ij}$  as given by (3.89). Application of the tangent law to the plane triangle  $CP_iP_j$  yields

$$\frac{(R_m + h_j - R_m - h_i)}{(R_m + h_j + R_m + h_i)} = \frac{\tan \frac{1}{2}(\pi - Z'_{ij} - \pi + Z'_{ji})}{\tan \frac{1}{2}(\pi - Z'_{ij} + \pi - Z'_{ji})}. \quad (16.82)$$

The realization that  $(Z'_{ij} + Z'_{ji}) - \pi = \psi_{ij} = S_{ij}^E/R_m$  and the expansion of the tan function in the denominator into a power series eventually yields

$$\Delta h_{ij} \doteq S_{ij}^E \left( 1 + \frac{h_m}{R_m} + \frac{S_{ij}^{E2}}{12 R_m^2} \right) \tan \left( \frac{Z'_{ji} - Z'_{ij}}{2} \right),$$

(16.83)

where  $h_m = \frac{1}{2}(h_i + h_j)$ .

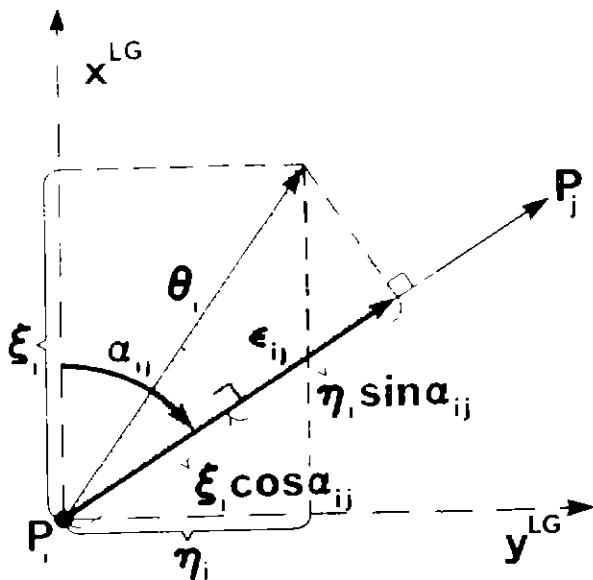


FIG. 16.21. Projected deflection of the vertical in desired direction  $\alpha_{ij}$ .

Unfortunately, because the observed zenith distances are usually close to  $\frac{1}{2}\pi$ , they are very sensitive to atmospheric refraction, as already seen in §15.2. This problem can be alleviated somewhat by making simultaneous zenith distance measurements. This procedure, under the best of conditions, results in an accuracy of  $\sigma_z = 1''$  and  $\sigma_{\Delta h} = 10$  cm for  $S^E = 10$  km [HEISKANEN AND MORITZ, 1967]. In practice, the accuracy achieved is usually significantly lower. The problem of atmospheric refraction will be treated more fully in the context of networks (§17.1). Let it suffice to state here that because of this problem, the trigonometrical height differences are not nearly as accurate as levelled height differences and, as such, are used neither in precise work nor in establishing height networks.

Next, let us study height difference determination by *geodetic levelling*. Levelling (for a description of the instrumentation and observing techniques needed see RAPPLEYE [1948]) does not suffer from as serious a refraction effect as zenith distance observations do. When the foresight  $\Delta S_F$  and backsight  $\Delta S_B$  (see FIG. 22) are balanced, the first-order effect, i.e., the one caused by a convex or concave path (depending predominantly on the sign of the vertical temperature gradient  $\Delta t$ ), cancels. We get  $\delta_B \sim \delta_F$ , and what remains is only the residual refraction effect, due to irregular air stratification, known to be much smaller; more about this effect will be said in §19.2.

Like trigonometrical determination, levelling is also affected by the earth's gravity field. To show this, let us consider a levelling line from point  $P_0$  at sea level to a point  $P$ , on top of a mountain (see FIG. 23). The equipotential surfaces are shown as being non-parallel (cf. §6.3). Let us assume now that the first levelling route goes up the left side and the second up the right side of the mountain. For each level setup, such as shown in FIG. 22, a height increment  $\delta h$  is obtained as the difference between backward and forward sightings on a level rod. The height of  $P$ , 'above sea level' (cf. §7.1) is then obtained by summing up the  $\delta h$  increments, first on the left

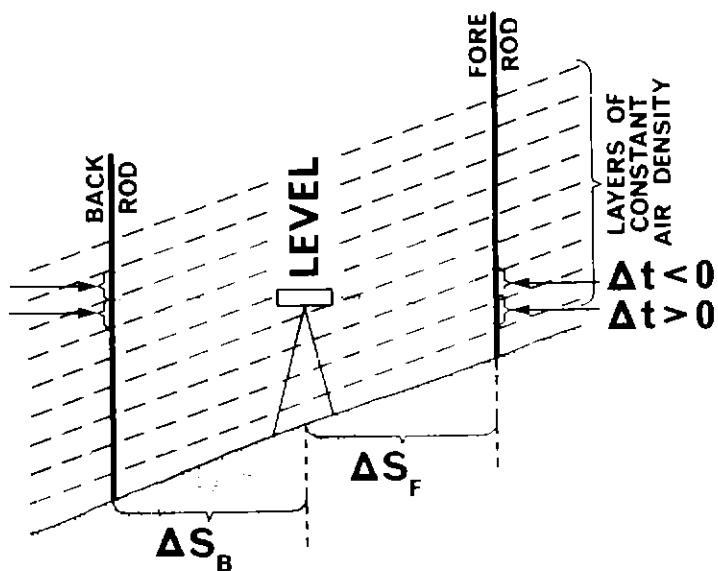


FIG. 16.22. Principle of geodetic levelling and character of refraction.

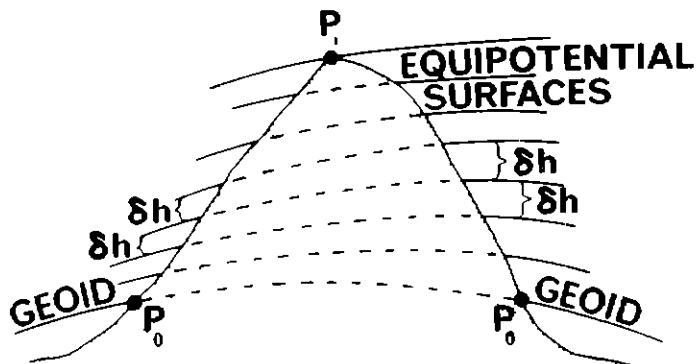


FIG. 16.23. Levelling.

side and then on the right side of the mountain. Clearly, two different values are obtained: there are greater separations  $\delta h$  between equipotential surfaces on the right side than on the left side. Which of the two values should be used for the height of  $P$ ?

The ambiguity can be eliminated only by converting the path-dependent results of levelling into unique path-independent height differences; this can be accomplished in a variety of ways. It is helpful to begin by showing how it can be done using the earth's gravity potential. According to (6.28), the difference between the potentials on two close together equipotential surfaces can be written as

$$\delta W = -g \delta h. \quad (16.84)$$

As explained in §6.3, only one equipotential surface passes through any point and, thus, there is only one value of potential  $W$  associated with each point. Therefore, the gravity potential represents one possible way of defining a unique vertical position. If the local spacing of equipotential surfaces (levelled height difference)  $\delta h$

is measured—from now on we shall denote the observed (levelled) height difference by  $\delta h$ —and the value of gravity  $g$  at the same location is known, the potential difference  $\delta W$  can then be evaluated from (84).

Instead of potential  $W_i$  of a point  $P_i$ , it is better to use the *geopotential number*  $C_i$  introduced by Tardi [BAESCHLIN, 1960]. It is defined as the negative potential difference between point  $P_i$  and the geoid:

$$C_i = -(W_i - W_0) = \int_{P_0}^{P_i} g \, dl = \int_{P_0'}^{P_i} g' \, dh', \quad (16.85)$$

where the integration is specified to proceed either along the terrain ( $dl$ ) between the geoid and the point  $P_i$  or along the plumb line ( $dh'$ ) of  $P_i$  (see FIG. 24). Similarly, the geopotential number difference  $\Delta C_{ij}$  between two points  $P_i$  and  $P_j$  is simply

$$\Delta C_{ij} = \int_{P_i}^{P_j} g \, dl. \quad (16.86)$$

The units for geopotential numbers adopted by the general assembly of the IAG at Rome in September 1954 are kilogal metres. The reason for this convention is that the numerical value of the geopotential number in these units will be approximately equal to the height  $H$  of the point above sea level in metres; more precisely,  $C = 0.98H$ .

Geopotential numbers have several useful properties: first of all, as stated above, they are unique for each point. Also, because the gravity potential is irrotational (see §6.3), the integral over a *closed loop*  $\mathcal{C}$  is zero, namely,

$$\oint_{\mathcal{C}} dC = \oint_{\mathcal{C}} g \, dl = 0. \quad (16.87)$$

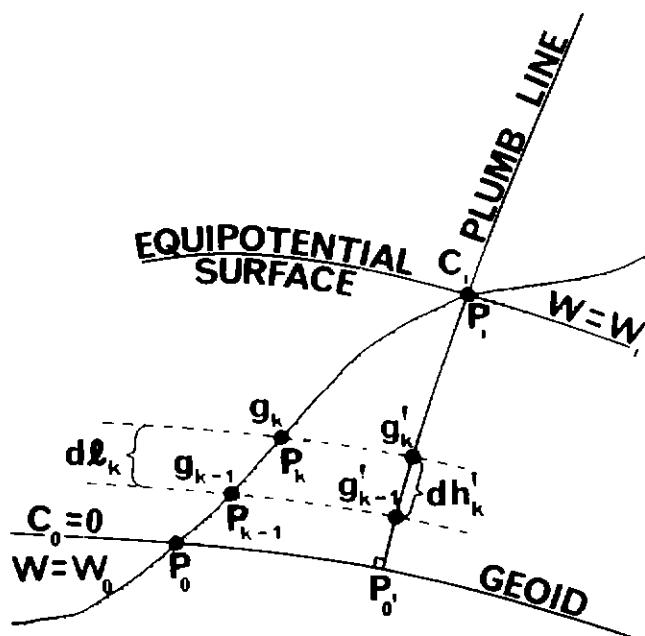


FIG. 16.24. Geopotential number.

This, evidently, is not true for levelled height differences. Further, the geopotential numbers are positive above the geoid, zero at the geoid, negative below it, and constant on the same equipotential surface. Finally, a geopotential number difference can be evaluated from observations made solely on the earth's surface.

In practice, neither  $l$  nor  $g$  is known as a continuous function of position. Therefore, the integrals in the above equations cannot be evaluated analytically, and it is necessary to resort to discretization employing measured values of  $g$  and  $\delta l$  along the levelled route. We have

$$\Delta C_{ij} \doteq \sum_{k=i}^j \bar{g}_k \delta l_k, \quad (16.88)$$

where

$$\bar{g}_k = \frac{1}{2}(g_{k-1} + g_k), \quad (16.89)$$

$\delta l_k$  is the observed level difference between adjacent bench marks, and  $g_k$  is the observed gravity value at the  $k$ th bench mark. From the practical point of view, it is neither feasible nor necessary to have the value of  $g$  observed at each bench mark. The only requirement is that  $g$ , whichever source it may be coming from, be known with sufficient accuracy. The errors due to inadequately spaced or inaccurate gravity values have been investigated by various researchers, e.g., HELMERT [1880] and LEVALLOIS [1964]. They found that the allowable spacing corresponding to the level differences  $\delta l$  and the accuracy of the gravity values are dictated by the nature of the terrain and the variability in the gravity field; the gravity effect will be discussed further in §19.2.

To eliminate the flaw of geopotential numbers not being expressed in length units, *dynamic heights*  $H^D$  have been introduced. They are obtained by dividing geopotential numbers by a constant reference gravity  $g_R$ , i.e.,

$$H_i^D = C_i / g_R. \quad (16.90)$$

Here,  $g_R$  can be thought of as the value of normal gravity on the mean earth ellipsoid (cf. §6.2) for a reference latitude  $\phi_R$  selected so that  $g_R$  represents approximately the average gravity in the region of interest. The reference gravity value can be viewed as the scale factor needed to convert the geopotential number in units of potential to units of length. The datum for dynamic heights is still the geoid. Nevertheless, one must be careful not to interpret the dynamic height of a point as the geometrical distance between the geoid and the point; in FIG. 25,  $l_i \neq l_1 \neq l_2 \neq l_3$ , but  $H_i^D = H_1^D = H_2^D = H_3^D$ . On the other hand, since the dynamic heights of points on one equipotential surface are equal, the dynamic heights are as useful as the geopotential numbers for various (such as hydrological) projects requiring some knowledge of the physical environment. It is even argued that either of these two height systems should be used in the design of super elevations for highway curves because they are physically akin to the accelerations considered there.

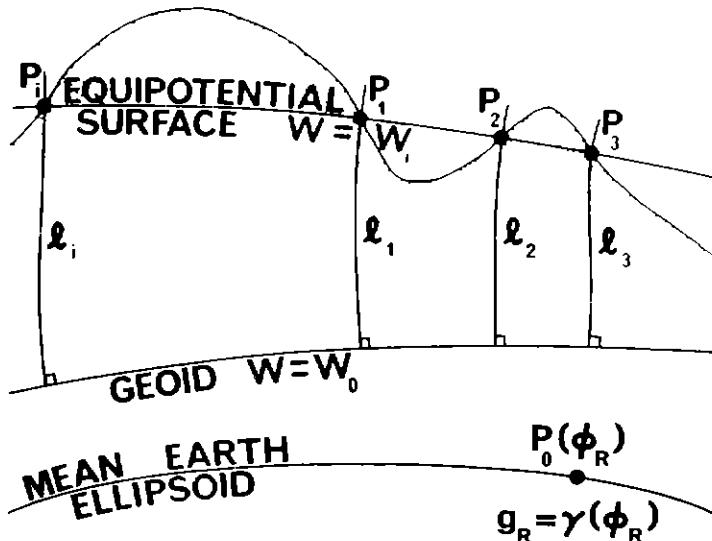


FIG. 16.25. Dynamic heights.

The dynamic height difference  $\Delta H_{ij}^D$  between two points  $P_i, P_j$  is defined as

$$\Delta H_{ij}^D = H_j^D - H_i^D = \frac{C_j}{g_R} - \frac{C_i}{g_R} = \frac{\Delta C_{ij}}{g_R}. \quad (16.91)$$

An alternative formula for the dynamic height difference is obtained by expressing it as a summation of the levelled height difference  $\Delta l_{ij}$  plus a correction, namely,

$$\Delta H_{ij}^D = \Delta l_{ij} + DC_{ij}. \quad (16.92)$$

The *dynamic correction*  $DC_{ij}$  is given by the obvious equation, which the reader may want to verify,

$$DC_{ij} = \sum_{k=i}^j \frac{\bar{g}_k - g_R}{g_R} \delta l_k. \quad (16.93)$$

Again, to evaluate this correction, it is not necessary to have gravity values observed at every bench mark; the criteria for the accuracy and spacing of gravity values are the same as in the case of geopotential numbers.

More intuitively appealing to many people is the geometrical concept of heights. The *orthometric height*  $H_i^O$  of a point  $P_i$  is defined as the geometrical distance between the geoid and the point, measured along the plumb line of  $P_i$  (see FIG. 26). The formula for  $H_i^O$  can be written simply as

$$H_i^O = \int_{P_0}^{P_i} dh', \quad (16.94)$$

where the integration is carried out along the plumb line. Substituting for  $dh'$  from (84) and denoting gravity along the plumb line by  $g'_i$  we get

$$H_i^O = - \int_{P_0}^{P_i} \frac{dW}{g'_i} = \int_{P_0}^{P_i} \frac{dC}{g'_i} = \int_{P_0}^{P_i} \frac{g}{g'_i} dt. \quad (16.95)$$

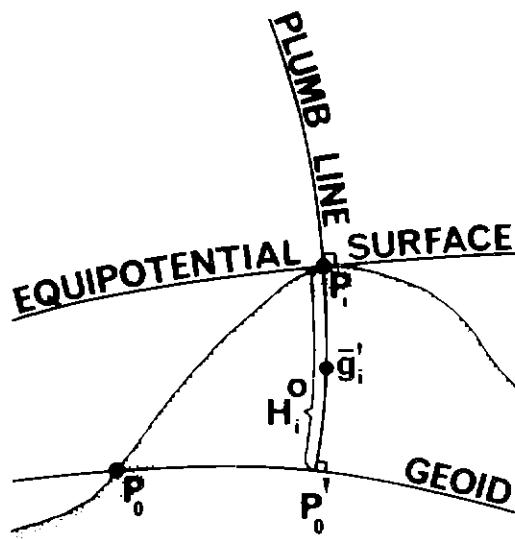


FIG. 16.26. Orthometric height.

The provenience of the second and third equations should be clear to the reader from the preceding discussion.

The formula for computing the orthometric height follows directly from the mean value theorem for integrals (see §3.2); in this context, it states that there exists a value of gravity  $\bar{g}'_i$  between the geoid and  $P_i$  such that

$$H_i^O = \frac{1}{\bar{g}'_i} \int_{P_0}^{P_i} g \, dl. \quad (16.96)$$

This  $\bar{g}'_i$  is the mean gravity along the plumb line of  $P_i$  in the integral sense. Then we can finally write

$$\boxed{H_i^O = C_i / \bar{g}'_i}, \quad (16.97)$$

where  $C_i$  is again the geopotential number of  $P_i$ . It is practically impossible to determine  $\bar{g}'_i$  because  $g$  along the plumb line is not known since the density distribution within the earth is not known. There are numerous approaches to approximating  $\bar{g}'_i$  each of which leads to a special kind of orthometric height usually ascribed to the proponent, e.g., to Niethammer, Mader, or Helmert. In all these approaches, an assumption must be made about the behaviour of the density within the earth; all these methods are thus only approximate.

Of all the proposed orthometric heights, the *Helmert orthometric height* is the one most often used in practice. It is defined as

$$H_i^H = C_i / g_i^H, \quad (16.98)$$

where the mean value  $g_i^H$  of gravity is taken as

$$g_i^H = g_i + 0.0424 H_i, \quad (16.99)$$

in which  $g_i$  is the gravity at  $P_i$  on the earth's surface. The numerical coefficient follows directly from the use of Poincaré–Pray's gravity gradient (see (21.38)) considered to be constant along the plumb line between the geoid and terrain; since the gravity gradient is considered constant,  $\bar{g}'$  is then evaluated directly for the midpoint of the plumb line of  $P_i$ . The physical unit of the second term is mgal for  $H_i$  in metres;  $H_i$  can be the observed height.

The major flaw of orthometric heights is that, for reasons stated earlier, they can never be determined exactly. This can also be understood to mean that the orthometric heights are never really referred to the geoid but to another reference surface more or less close to the geoid, depending on the kind of  $\bar{g}'$  used in the particular definition. However, when  $H^O$  goes to zero, so does the difference between the implied reference surface and the geoid which means that they coincide at least on the open sea. Also, because of the definition of orthometric heights, points on the same equipotential surface (except on the geoid) do not generally have the same orthometric heights; water may flow up hill.

Since it will never be theoretically possible to come up with proper orthometric heights, in 1954 MOLODENSKIY ET AL. [1960] suggested that they be replaced with *normal heights*  $H^N$ . These are defined as

$$H_i^N = C_i / \bar{g}_i, \quad (16.100)$$

where  $\bar{g}_i$  is the normal counterpart of  $\bar{g}'_i$ , i.e., the mean normal gravity along the plumb line of  $P_i$ . The difference is that while the mean actual gravity  $\bar{g}'_i$  was computed from the downward continuation of the actual surface gravity  $g_i$ , the mean normal gravity  $\bar{g}_i$  is computed from the upward continuation of normal gravity  $\gamma_0$  on the geocentric reference ellipsoid. While the mean actual gravity is sought on the actual plumb line between points of heights  $H$  (terrain) and 0 (geoid), mean normal gravity is sought on the normal plumb line between points of height 0 (geocentric reference ellipsoid) and  $H$  (telluroid)—cf. §7.4. The situation is shown in FIG. 27.

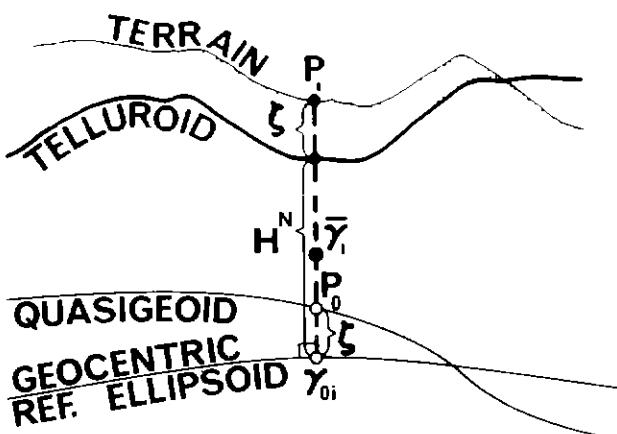


FIG. 16.27. Normal height.

To get  $\bar{\gamma}_i$  from  $\gamma_{0i}$ , Molodenskij uses an accurate formula (see (21.30)) for the vertical gradient of normal gravity. He has also shown [MOLODENSKIJ ET AL., 1960] that normal heights defined in this way can be regarded as referring to the quasigeoid (cf. §7.4) and as being measured along a normal plumb line. Hence normal heights can be, for all practical purposes, viewed as a special kind of orthometric heights referred to a reference surface (quasigeoid) that, compared with the reference surfaces of all the orthometric systems, happens to be determinable, as will be shown in §22.2.

Vignal [VIGNAL AND KUKKAMÄKI, 1954] proposed a similar system of heights, whereby the mean normal gravity is evaluated from  $\gamma_{0i}$  by means of approximate vertical gradient of normal gravity (see (21.32)), which happens to coincide numerically with the free-air gravity gradient (see (6.14)). Then the *Vignal normal height* is given by

$$H_i^V = \frac{C_i}{\gamma_{0i} - 0.1543 H_i}. \quad (16.101)$$

The Vignal and Molodenskij heights are numerically quite close. The Vignal height has been adopted for the unification of the European levelling networks [SIMONSEN, 1963], while the normal height of Molodenskij is used in the U.S.S.R. and the east European countries. The Vignal height has also been proposed for use in North America [KRAKIWSKY AND MUELLER, 1966].

The problem common to all three height systems, i.e., dynamic, orthometric, and normal, used to be, and in some countries still is, the lack of available actual gravity data on the earth's surface necessary for the evaluation of the appropriate corrections. This problem was by-passed by using normal gravity  $\gamma(\phi, H)$  instead of actual  $g$ . *Height differences* (dynamic, orthometric, or normal) *based on normal gravity* determined in this way do not depart excessively from the proper values when only adjacent bench marks are considered. The effect of this approximation in networks of points is, however, more serious and will be dealt with in §19.2.

*Height differences determined by three-dimensional methods* (§16.1) are obtained as follows: first, the relative position between two points is expressed in the G system. This is then transformed to differences in  $\phi, \lambda, h$  using one of the two techniques described in §15.4; the height difference  $\Delta h_{ij}$  with respect to the reference ellipsoid is one of the three results. In this approach, one can use the classical terrestrial, extraterrestrial, three-dimensional methods, as well as inertial positioning results. Of specific interest in this context is the method of *Doppler levelling* [KOUBA, 1976] in which the concept of translocation (§16.1) is used, while the solution is restricted to the determination of geodetic height differences (above the ellipsoid). If the geoid height difference  $\Delta N_{ij}$  is known, then the orthometric height difference  $\Delta H_{ij}^O$  can also be determined. The accuracy of the height difference that can be obtained by this method is of the order of 1 m.

## CHAPTER 17

### THREE-DIMENSIONAL NETWORKS

Three-dimensional networks have already been defined and introduced in §7.1. They fall into three natural classes: networks established using the standard terrestrial observables (horizontal angles and distances) and astronomical quantities; photogrammetrical networks utilizing photographs taken from the air; and networks based on observations made from tracking stations to orbiting satellites. These three classes of networks are treated respectively in the first three sections.

The problem of a rigorous definition of the position and orientation of the coordinate system, used to describe the network mathematically, is not a trivial one. Usually a geodetic (G) coordinate system is used, and its position and orientation with respect to the earth has to be fixed. Some of the ways this can be done have been shown in §15.4; other ways will be explained here in the appropriate context. An understanding of the concept of positioning and orienting a coordinate system is particularly needed within the context of merging different three-dimensional networks that have been established by independent means. This task is dealt with in the fourth section where the ideas related to the design and assessment of three-dimensional networks are also set forth.

When dealing with the material of this chapter, the earth is regarded as rigid and no allowance is made for its deformations in time; the determination of these deformations is treated separately in Part VI.

#### 17.1. Three-dimensional networks using terrestrial observations

In §16.1 it was shown how the relative position of one point with respect to another can be computed in three dimensions from the standard geodetic observables of spatial distances, zenith distances, and astronomical latitudes, longitudes, and azimuths. In this section, we extend the technique to a whole network of points and allow for the inclusion of two additional observables: horizontal directions or, alternatively, horizontal angles, and height differences. In contrast to the computations in two dimensions, on both the ellipsoid and the conformal mapping plane (cf. §16.2, and §16.3) no reductions of the observables are necessary here except for instrument errors and refraction. As much as possible, instrument errors are eliminated through measurement procedures; the bulk of the refraction effects is

usually corrected for by means of modelling, and the residual refraction may be parameterized and eliminated at the solution stage. The theoretical foundations for modelling three-dimensional networks have been laid by BRUNS [1878], HOTINE [1969], and others.

The basic unit of the mathematical model needed in the *adjustment of a three-dimensional network* (cf. §14.3) is the interstation vector defined in §16.1. The interstation vectors for all pairs of adjacent points  $P_i, P_j$  constitute the network. The mathematical formula for an interstation vector in the LA system has already been given by (16.1). In a general form, it can be written as

$$\underbrace{f_{ij}(x_i, y_i, z_i, x_j, y_j, z_j; A_{ij}, \nu_{ij}, \Delta r_{ij}, \Phi_i, \Lambda_i)}_{\text{unknowns}} = 0. \quad (17.1)$$

The Cartesian coordinates of points  $P_i$  and  $P_j$  are the unknown parameters to be determined, and the observables are the astronomical latitude  $\Phi_i$ , longitude  $\Lambda_i$ , and azimuth  $A_{ij}$ , vertical angle  $\nu_{ij}$  (or, equivalently, the zenith distance  $Z_{ij}$ ), and the spatial distance  $\Delta r_{ij}$ . Clearly,  $\Phi_i$  and  $\Lambda_i$  define the direction of gravity at point  $P_i$  and thus serve as a reference direction in space to which  $A_{ij}$  and  $\nu_{ij}$  (or  $Z_{ij}$ ) are then referred. For the inverse interstation vector (from  $P_j$  to  $P_i$ ), the measurements ( $A_{ji}$ ,  $\nu_{ji}$  (or  $Z_{ji}$ ),  $\Phi_j$ ,  $\Lambda_j$ ,  $\Delta r_{ji} = \Delta r_{ij}$ ) are made at the other end of the line, and the subscripts in (1) are interchanged.

The above models are implicit. For practical reasons, a model explicit in observables (cf. §10.1) is preferred, so that an omission of a measurement leads simply to a deletion of an observation equation without any further consequences. VINCENTY [1973] has formulated the explicit model in both Cartesian  $(x, y, z)^G$  and curvilinear  $(\phi, \lambda, h)$  coordinates. This explicit model also includes horizontal directions  $d$  or angles  $\omega$  as well as height differences  $\Delta h$ . The model is then linearized and the solution sought in the form of, for example, corrections  $\delta\phi$ ,  $\delta\lambda$ , and  $\delta h$  to the approximate values  $\phi^{(0)}$ ,  $\lambda^{(0)}$ , and  $h^{(0)}$  of the unknowns. These approximate values are obtained through repeated relative positioning, as described in §16.1, applied from point to point or any other equivalent technique.

The linearized observation equations that make up the model are obtained by linearization from (16.1), transformed into the  $(\phi, \lambda, h)$  coordinate system (see §12.2). The *astronomical azimuth observation equation* for the interstation vector  $\Delta \vec{r}_{ij}$  reads as follows:

$$\boxed{r_{ij}^A = a_1 \delta\phi_i + a_2 \delta\lambda_i + a_3 \delta h_i + a_4 \delta\phi_j + a_5 \delta\lambda_j + a_6 \delta h_j + a_7 \delta\Phi_i + a_8 \delta\Lambda_i + \Gamma_i^{(0)} - 4} \quad (17.2)$$

The coefficients,  $a_i$ ,  $i = 1, \dots, 8$ , are functions of both the unknowns and observables and are listed in TABLE I (together with the coefficients for the other kinds of observation equations that make up the design matrix). The expansion value of  $A_{ij}$  used for the linearization is denoted by  $A_{ij}^{(0)}$ ;  $r_{ij}^A$  is the residual of  $A_{ij}$ . Note that the astronomical coordinates are also treated as unknown parameters.

TABLE 17.1  
Coefficients of design matrix for three-dimensional networks (according to VINCENTY [1973]). (Prime denotes quantities related to the second point,  $P_j$ )

Un-known	sub-script	Observed $A$ or $d$ (or $\omega$ )	Observed $\nu$ (or $Z$ )		Observed $\Delta r$
			$a$	$b$	
$\Phi_i$	1	$(M + h)\sin A / (\Delta r \cos \nu)$	$(M + h)\cos A \sin \nu / \Delta r$	$-(M + h)\cos A \cos \nu$	
$\lambda_i$	2	$-(N + h)\cos \phi \cos A / (\Delta r \cos \nu)$	$(N + h)\cos \phi \sin A \sin \nu / \Delta r$	$-(N + h)\cos \phi \sin A \cos \nu$	
$h_j$	3	0	$-\cos \nu / \Delta r$	$-\sin \nu$	
$\Phi_j$	4	$-(M' + h')\sin \phi' \cos \Delta \lambda \sin A$ $+ \sin \phi' \cos A \sin \Delta \lambda + \cos \phi' \cos \phi' \sin A)$ $(\Delta r \cos \nu) \doteq (M' + h')\sin A' / (\Delta r \cos \nu')$	$-(M' + h')(\cos \phi' \sin \phi' \cos \Delta \lambda - \sin \phi' \cos \phi' - \cos A' / \times \sin \nu \cos \nu') \sec \nu / \Delta r$	$-(M' + h')\cos A' \cos \nu'$ $\times \sec \nu / \Delta r \doteq -(M' + h')\cos A' \sin \nu' / \Delta r$	
$\lambda_j$	5	$(N' + h')\cos \phi' (\cos \Delta \lambda \cos A - \sin \phi' \sin \Delta \lambda \sin A) / (\Delta r \cos \nu) \doteq -(N' + h')\cos \phi' \cos A' / (\Delta r \cos \nu')$	$-(N' + h')\cos \phi' (\cos \phi' \sin \Delta \lambda - \sin A' \sin \nu \cos \nu') / \times \sec \nu / \Delta r \doteq -(N' + h')\cos \phi' \sin A' \sin \nu' / \Delta r$	$-(N' + h')\cos \phi' \sin A' \cos \nu'$ $\times \sec \nu / \Delta r \doteq -(N' + h')\cos \phi' \sin A' \sin \nu' / \Delta r$	
$h_j$	6	0 (for $ h_j - h_i  < 1$ m)	$(\cos \phi' \cos \phi' \cos \Delta \lambda + \sin \phi' \sin \phi' + \sin \nu \sin \nu') \times \sec \nu / \Delta r \doteq \cos \nu' / \Delta r$	$-\sin \nu'$ $\times \sec \nu / \Delta r \doteq \cos \nu' / \Delta r$	
$\Phi_i$	7	$\sin A \tan \nu$	$\cos A$	0	
$\Lambda_j$	8	$\sin \phi - \cos \phi \cos A \tan \nu$	$\cos \phi \sin A$	0	

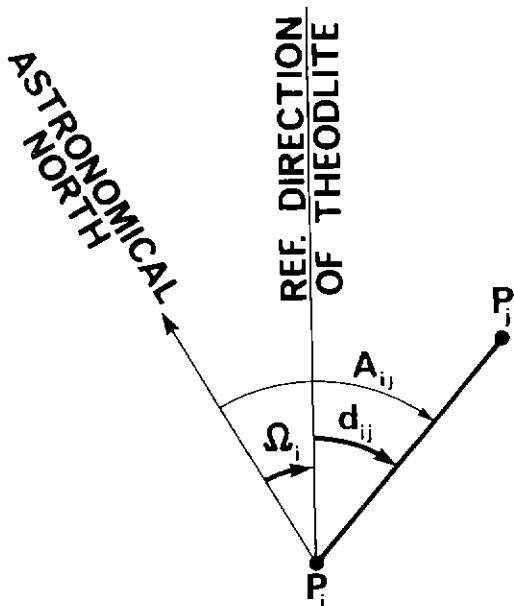


FIG. 17.1. Orientation of horizontal directions.

The *direction observation equation* follows from the above simply by replacing  $A_{ij}$  with  $d_{ij} + \Omega_i$  (see FIG. 1) where  $\Omega_i$  is the *orientation unknown* at  $P_i$ , and  $d_{ij}$  is the direction measured with a theodolite referred to an arbitrary reference. We get

$$\boxed{r_{ij}^d = a_1 \delta\phi_i + a_2 \delta\lambda_i + a_3 \delta h_i + a_4 \delta\phi_j + a_5 \delta\lambda_j + a_6 \delta h_j + a_7 \delta\Phi_i + a_8 \delta\Lambda_i - \delta\Omega_i + A_{ij}^{(0)} - d_{ij} - \Omega_i^{(0)}}, \quad (17.3)$$

where  $A_{ij}^{(0)}$  is the value of  $A_{ij}$  used in linearization. The addition of  $\Omega_i$  to the observed direction transforms the theodolite's reference to astronomical north; the approximate value of the orientation unknown,  $\Omega_i^{(0)}$ , is the computed approximate geodetic azimuth of this reference direction. At least one separate orientation unknown must appear at each station where directions have been measured.

A horizontal angle  $\omega_{ijk}$  can be considered as being simply the difference between two directions,  $d_{ik}$  and  $d_{ij}$ —see FIG. 2. The *horizontal angle observation equation* is thus obtained by differencing two direction observation equations referring to lines  $P_iP_j$  and  $P_iP_k$ :

$$\boxed{r_{ijk}^\omega = (a_1(k) - a_1(j)) \delta\phi_i + (a_2(k) - a_2(j)) \delta\lambda_i + (a_3(k) - a_3(j)) \delta h_i + a_4(j) \delta\phi_j + a_5(j) \delta\lambda_j + a_6(j) \delta h_j + a_4(k) \delta\phi_k + a_5(k) \delta\lambda_k + a_6(k) \delta h_k + (a_7(k) - a_7(j)) \delta\Phi_i + (a_8(k) - a_8(j)) \delta\Lambda_i + \omega_{ijk}^{(0)} - \omega_{ijk}}, \quad (17.4)$$

The *vertical angle observation equation* reads

$$\boxed{r_{ij}^v = b_1 \delta\phi_i + b_2 \delta\lambda_i + b_3 \delta h_i + b_4 \delta\phi_j + b_5 \delta\lambda_j + b_6 \delta h_j + b_7 \delta\Phi_i + b_8 \delta\Lambda_i - \delta\nu + v_{ij}^{(0)} - v_{ij}}, \quad (17.5)$$

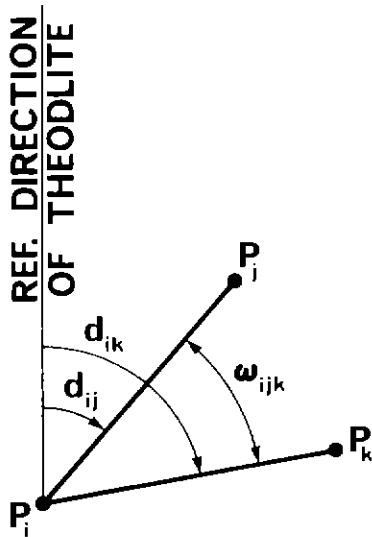


FIG. 17.2. Horizontal angle.

where  $\delta\nu$  stands for the unknown residual vertical angle refraction correction, known as the *residual refraction correction*, to be discussed in detail below. For the formulae for  $b_i$ ,  $i = 1, \dots, 8$ , see TABLE 1. The *spatial distance observation equation* reads

$$r_{ij}^{\Delta r} = c_1 \delta\phi_i + c_2 \delta\lambda_i + c_3 \delta h_i + c_4 \delta\phi_j + c_5 \delta\lambda_j + c_6 \delta h_j + \Delta r_{ij}^{(0)} - \Delta r_{ij}, \quad (17.6)$$

where again the formulae for the coefficients are given in TABLE 1.

The units of all the coefficients are such that corrections to angular quantities, including the orientation unknown and residual refraction correction, are in radians. In practice, it is usual to convert these to seconds of arc. According to VINCENTY [1973], if the approximate coordinates  $\phi^{(0)}, \lambda^{(0)}, h^{(0)}$  of points in the network are known to about 1 m, the coefficients  $a_4, a_5, b_4, b_5$ , and  $b_6$  to be used are those shown in TABLE 1. If, e.g., after some iterations, the approximate values become known to 0.1 m or better, then we can set  $M + h \doteq N + h \doteq R$  in all the coefficients. The covariance matrix  $C$  of the observations is assembled simply from the variances and covariances of the individual observations. For the variances, it is advisable to choose the same physical units (squared) as those of the absolute terms of the model so that the residuals, computed from the model, are also in these units.

If the astronomical coordinates  $\Phi, \Lambda$  have also been observed, two more kinds of observation equations, called *astronomical coordinate observation equations*, can be added to the linearized model: namely,

$$r_i^\phi = \delta\Phi_i + \Phi_i^{(0)} - \Phi_i, \quad (17.7)$$

$$r_i^\lambda = \delta\Lambda_i + \Lambda_i^{(0)} - \Lambda_i. \quad (17.8)$$

If, instead of the astronomical coordinates, the surface deflection components  $\xi, \eta$  are available, a similar system of observation equations can be written for them. The

*deflection component observation equations* read

$$r_i^\xi = \delta\Phi_i - \delta\phi_i, \quad r_i^\eta = \sec\phi_i \delta\Lambda_i - \sec\phi_i \delta\lambda_i. \quad (17.9)$$

As mentioned earlier, it is also possible to include measured height differences  $\Delta h$  in three-dimensional networks. This type of information may come from spirit levelling or from an extraterrestrial technique, as shown in §16.4. The *height difference observation equation* is

$$r_{ij}^{\Delta h} = -\delta h_i + \delta h_j + \Delta h_{ij}^{(0)} - \Delta h_{ij}, \quad (17.10)$$

where all the heights are referred to a specific reference ellipsoid (cf. §16.4). It is also possible to deal directly with orthometric heights (see, e.g., CHOVITZ [1974]), but in this case the geoidal heights have to be accounted for. Note that this procedure can be followed in merging horizontal with height networks into three-dimensional networks, as discussed in §7.1. It should be realized that not all the observables in the network have to be observed. On the other hand, some of the observables must be observed to prevent the model from becoming singular (cf. §14.5). The selection of necessary observations is done on the basis of the network geometry. A simple case of three points in two dimensions will be treated in §18.4.

When all observables, including the astronomical coordinates, are measured, the complete model can be regarded as equivalent to the combination of two models (cf. §14.4); namely,

$$f_1(x_1, x_2, t_1) = 0, \quad (17.11)$$

$$f_2(x_2, t_2) = 0, \quad (17.12)$$

where we have denoted  $x_1 = [\phi, \lambda, h]^T$ ,  $x_2 = [\Phi, \Lambda]^T$ ,  $t_1 = [A, \nu, \Delta r]^T$ , and  $t_2 = [\Phi, \Lambda]^T$ . The astronomical coordinates are treated as unknown parameters in the first model and as both directly observed and unknown in the second. Thus, it is easy to see what happens when  $\Phi$  and  $\Lambda$  are not measured:  $f_2$  is simply deleted and the unknown  $\delta\Phi$  and  $\delta\Lambda$  are solved for from  $f_1$ . A similar argument holds as well for observed height differences.

FUBARA [1972] and others have found that the height above the ellipsoid can be computed with about the same accuracy as can the horizontal coordinates (geodetic latitude and longitude) provided rigorous mathematical models are used and proper measurement procedures followed. Until this finding, the vertical component had always been questioned because of the fear of refraction effects.

A more detailed discussion of the refraction effect on  $\nu$  (or  $Z$ ) is now in order. There are basically three options available for handling the effect:

(a) The first is simply to correct the observed  $\nu$  for the whole effect, and delete the unknown  $\delta\nu$  from the model.

(b) The second is to correct the observations for the first-order effect using a simpler refraction model, and leave the unknown  $\delta\nu$  in to absorb the residual refraction (see FIG. 3). The first-order effect  $\Delta\nu$  can be evaluated more reliably if simultaneous measurement of the vertical angles (zenith distances) at each end of the

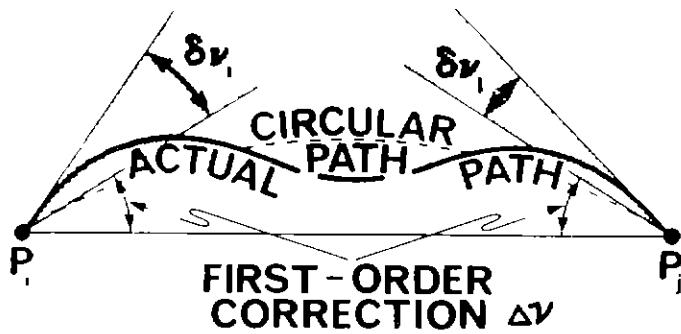


FIG. 17.3. First-order and residual refraction correction.

line are available. From FIG. 16.20, we can write

$$\Delta\nu_{ij} \equiv \Delta\nu_{ji} = \frac{1}{2}\pi - \frac{1}{2}(Z_{ij} + Z_{ji} - \psi), \quad (17.13)$$

where the two zenith distances have been corrected for the effect of the deflection of the vertical (see (16.81)).

(c) The last option entails a different measurement technique by which redundant vertical angles from the same  $P_i$  to several adjacent points are measured in rapid succession. Details about this option, as well as other intricacies of vertical angle measurements in three-dimensional networks, can be found in, e.g., HRADÍLEK [1972].

At this juncture, it is worthwhile to point out that it is possible to assemble a three-dimensional network from a horizontal network (of known geodetic horizontal coordinates  $\phi, \lambda$ ) and a height network (of known orthometric  $H^O$  or normal heights  $H^N$ ). The requirements are that the points of the two networks be common—generally not the case (cf. §7.1) unless a special effort is made when the networks are designed—and the geoidal height  $N$  (or height anomaly  $\xi$ ), referred to the same reference ellipsoid as  $\phi$  and  $\lambda$ , is known at each point. The points of the assembled three-dimensional network are then described by triplets of geodetic coordinates  $(\phi, \lambda, h)$ , where  $h = H^O + N = H^N + \xi$ .

How is the coordinate system, to which the three-dimensional network is referred, positioned and oriented with respect to the earth? The explicit positioning and orientation of this coordinate system is done either geocentrically by specifying the translation vector and three rotation angles with respect to another coordinate system, usually the CT system (see §15.4(b)), or topocentrically by specifying six parameters (see §15.4(c)) at one point of the network. Such a point is called the *origin of the network*, or the *initial point*, and the selected six parameters are held fixed in the adjustment of the network. The explicit positioning is discussed in detail by HOTINE [1969].

In the implicit positioning and orientation of the coordinate system, the position and orientation are implied by the coordinates of either some or all the points within the network [KOLACZEK AND WEIFFENBACH, 1975]. The coordinates of these defining points are arrived at either from some prior geodetic work or from an adjustment of a singular model using a generalized matrix inverse or inner con-

straints (see §14.5). The constraint matrix  $D_i$  needed for the inner constraint adjustment in Cartesian coordinates  $x, y, z$  reads [BLAHA, 1971a]

$$D_i = \begin{bmatrix} 1 & 0 & 0 & | & 1 & 0 & 0 & | & 1 & 0 & 0 \\ 0 & 1 & 0 & | & 0 & 1 & 0 & | & 0 & 1 & 0 \\ 0 & 0 & 1 & | & 0 & 0 & 1 & | & 0 & 0 & 1 \\ -z_1 & -y_1 & 0 & | & 0 & z_2 & -y_2 & | & 0 & z_n & -y_n \\ -z_1 & 0 & x_1 & | & -z_2 & 0 & x_2 & | & -z_n & 0 & x_n \\ y_1 & -x_1 & 0 & | & y_2 & -x_2 & 0 & | & y_n & -x_n & 0 \end{bmatrix}, \quad (17.14)$$

where  $n$  is the number of points in the network. A more detailed treatment of implicit positioning (of a reference ellipsoid) will be given in §18.1.

## 17.2. Photogrammetrical networks

For the purpose of densification of networks (see §7.1), photogrammetrical methods have long been utilized. A *photogrammetrical block* is composed of partially overlapping aerial photographs containing some of the points belonging to the network to be densified, usually in this context called control points. Positions of these control points known in a coordinate system provide the necessary information to give the positions of other points visible on the photographs in the same coordinate system.

Considering the observational information in the form of photocordinates  $(x, y)^P$ , i.e., coordinates of the desired points and known control points measured on the photographic plates, a rigorous model can be formulated. Such a model usually explicitly includes the transformation parameters between the photographic plate coordinate system and the coordinate system of control points. The inclusion of auxiliary information to constrain the adjustment, or the inclusion of some calibration parameters to refine the formulation, may be considered in some cases.

A photographic image of a (three-dimensional) patch of the earth's surface is simply the result of a central projection of that patch onto the photographic plate through the *perspective centre*. Mathematically, the photocordinates  $(x, y)^P$  are related to the ground coordinates, say,  $(x, y, z)^G$ , in the following manner:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}^G = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}^P, \quad (17.15)$$

where the six coefficients  $a_{ij}$  correspond to the six degrees of freedom of such a projection. The six degrees of freedom of a projection in object space can be visualized in terms of the position of the perspective centre,  $\tilde{r}_0^G = (x_0, y_0, z_0)^G$ , and the orientation  $\kappa$  of the bundle of rays about the principal axis of a specific direction

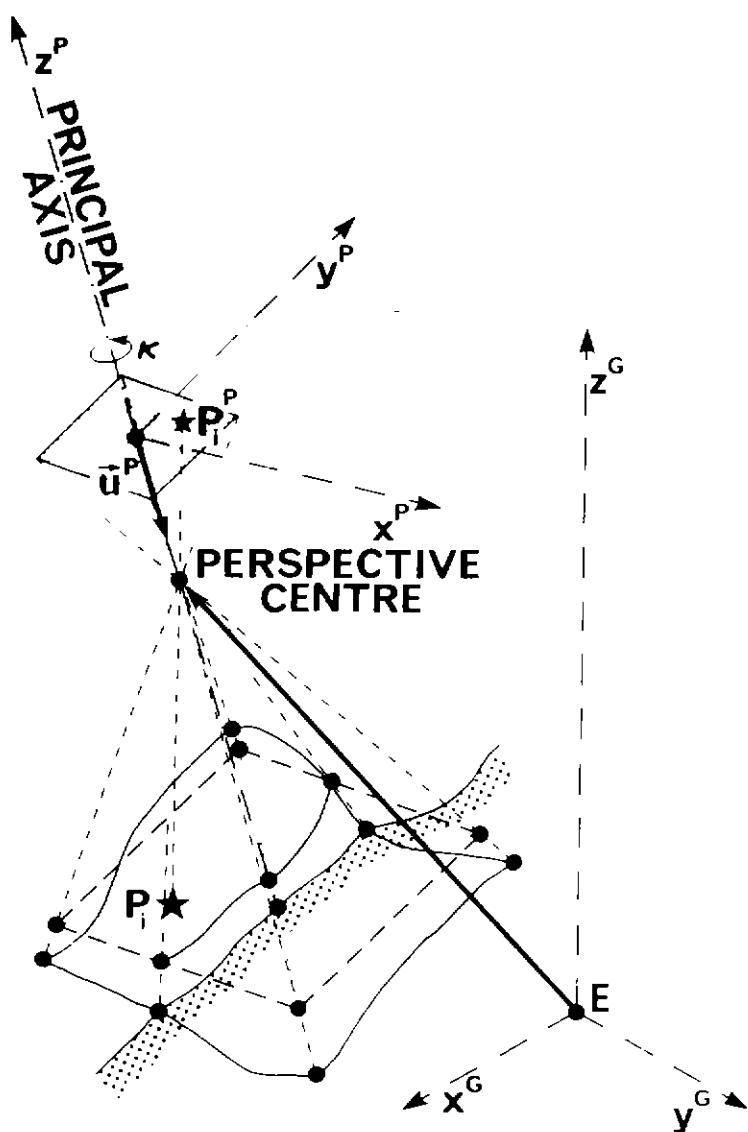


FIG. 17.4. Geometry of a single aerial photograph.

given by the unit vector (see FIG. 4)

$$\vec{e}^P = \begin{bmatrix} \cos \mu \cos \beta \\ \cos \mu \sin \beta \\ \sin \mu \end{bmatrix} \quad (17.16)$$

A model can then be formulated to include the six parameters in each photograph, or bundle. The model may also include a few nuisance parameters associated with the focal length of the camera, the definition of the coordinate system  $(x, y)^P$  relative to the camera [WOLF, 1974], and the photocordinates as well as geodetic coordinates of the control points.

The other approach is based on the idea of reconstructing the three-dimensional model, called the *stereomodel* (FIG. 5), of the photographed patch from a pair of

## STEREOPHOTOGRAPHS

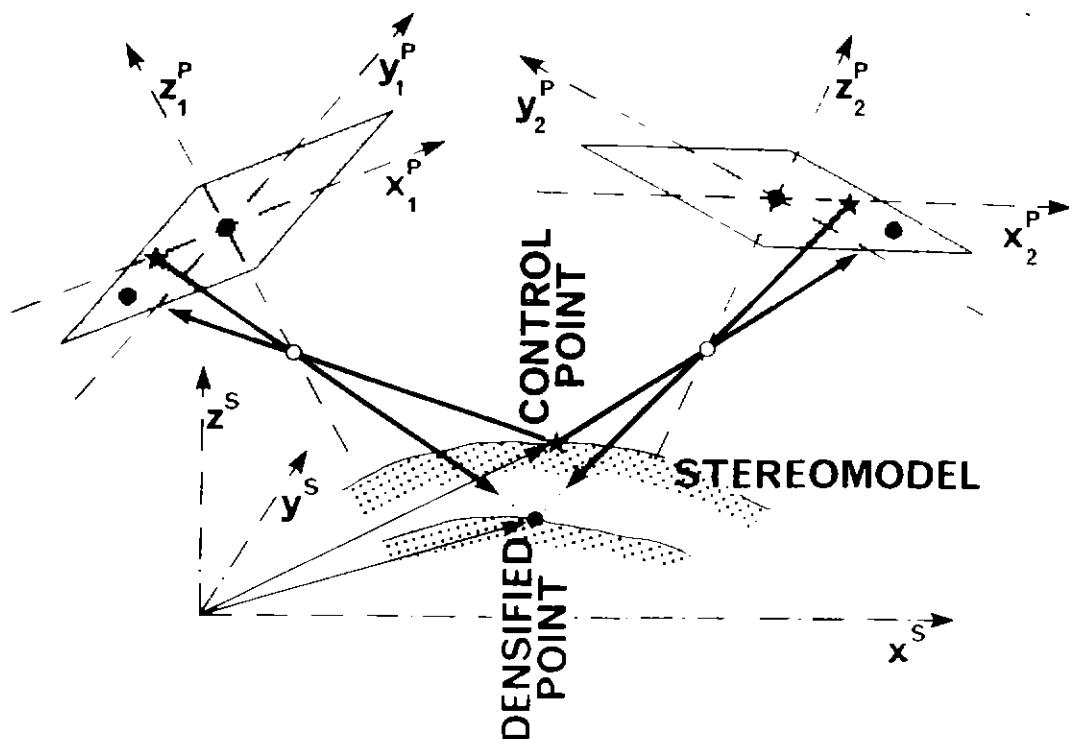


FIG. 17.5. Stereomodel.

suitably projected photographs (stereophotographs). The relation between a stereomodel space and an object space is a spatial similarity, as the stereomodel is an exact reconstruction, except for scale, orientation, and location, of the situation in which the photographs were taken. A stereomodel, taken as an isolated entity, has seven degrees of freedom: three rotations, three translation components, and one scale factor. These correspond to the seven degrees of freedom of a *similarity transformation* in three-dimensional space. A linearized similarity transformation between the stereomodel coordinate system (S) and the corresponding control point coordinate system, say, the G system, can be simply written as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}^G = \begin{bmatrix} a & b & -c \\ -b & a & d \\ c & -d & a \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}^S + \begin{bmatrix} e \\ f \\ g \end{bmatrix}^G, \quad (17.17)$$

in which  $a, b, c, d, e, f$ , and  $g$  are equivalent to the seven parameters above. The adjustment of the photogrammetrical model can be carried out either directly in one step (a) or in phases (b).

(a) The direct approach is usually called *bundle block adjustment*, because it solves explicitly for the projective bundles of rays corresponding to the aerial photographs. Since about the 1960s, a number of different mathematical models and corresponding computer programs for the bundle block adjustment have been developed by, e.g., KELLER [1967] and SCHUT [1968].

(b) The phased approach is known as the *stereomodel block adjustment*. The adjustment usually proceeds in three phases:

- A stereomodel formation phase consisting of the algebraic elimination of all the nuisance parameters pertaining to the orientation of the individual photographs needed to create the stereomodel.
- A block adjustment phase consisting of the adjustment of the resulting stereomodel, including all the correlations, to fit the terrestrial control points.
- A back substitution phase consisting of an evaluation of the parameters eliminated in the first phase through the use of the adjusted values of transformation parameters obtained in the second phase.

As we have seen in §14.6, the phased approach is rigorously equivalent to the simultaneous approach as long as the same observational information is used in both approaches and the covariance matrices are propagated properly. Another point worth noting is that the phase approach with stereomodels can be used in a variety of ways, as one could work, for instance, with triplets, quadruplets, or other groupings of stereomodels. The choice has to be based on other considerations, such as instrumentation, data processing, personnel, and so on.

The stereomodels can also be formed using analogue or semi-analogue procedures instead of the aforementioned analytical ones. In such a case, some observational information is often left out as are the appropriate covariance matrices. On the other hand, the expediency of the procedures used to eliminate certain nuisance parameters in the stereomodel compensates for some of the negative effects. In fact, several experiments with actual photography have indicated that, under certain conditions, this *semi-analytical stereomodel adjustment* is more advantageous than the bundle adjustment (e.g., ACKERMANN [1974]). Generally the gap between analytical and analogue techniques has been somewhat bridged by the introduction of self-calibration bundle adjustment methods that account for systematic effects and correlations (e.g., EBNER [1975]).

The least-squares fitting of blocks of stereomodels to terrestrial control points can be done either in horizontal coordinates with levelled models (e.g., VAN DEN HOUT [1966]), or sequentially in horizontal coordinates and height (e.g., ACKERMANN [1968]), or directly in three-dimensional space (e.g., BLAIS [1979]). The number of unknown similarity transformation parameters for each model are respectively four, four plus three, and seven. The computational effort required increases greatly with the number of unknown parameters for each model, especially when the adjustment is done in a simultaneous manner.

Once the adjustment of either the bundles or the stereomodels is done, geodetic coordinates of points required for the densification are obtained very simply. These points are first identified on the photographs or models and their coordinates  $(x, y)^P$  or  $(x, y, z)^S$  are then measured (cf. FIG. 5). Then application of either (15) or (17) yields the desired result, i.e.,  $(x, y, z)^G$ .

The propagation of random errors in photogrammetrical networks is shown in FIG. 6. The accuracy of the photogrammetrically determined horizontal coordinates within a block consisting of numerous stereomodels surrounded by dense perimeter control (whose accuracy is denoted by  $\sigma_g$ ) is represented by  $\sigma$  — the major semi-axis

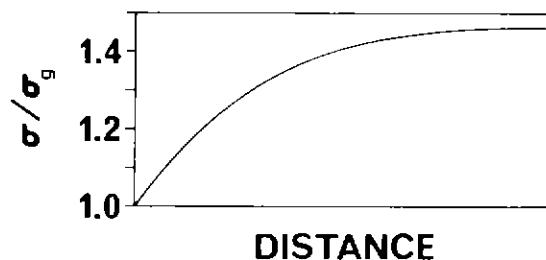


FIG. 17.6. Propagation of random errors in a photogrammetrical network with distance from perimeter control [EBNER, 1975].

of the standard point error ellipse. It can be seen that  $\sigma$  increases logarithmically as one moves away from the perimeter control, approaching asymptotically  $1.5 \sigma_g$ .

### 17.3. Three-dimensional networks using extraterrestrial observations

Satellite point positioning (§15.3) and satellite relative positioning (§16.1) involved only one or two tracking stations respectively. With these simple configurations, it was not possible to collect data with the appropriate distribution needed to model exhaustively the orbital biases and systematic effects arising from the measurement process. With a network of points, however, it is possible both to refine the orbits and to model systematic effects successfully, resulting then in more accurate coordinates for the points in the network. The ensuing discussion will be based on the use of satellites, but it should be understood that any extraterrestrial object may be employed instead.

The two modes used for establishing three-dimensional networks using extraterrestrial observations are the geometrical mode (a) and the kinematically constrained mode (b).

(a) Let us begin by taking a closer look at the *geometrical mode of simultaneous positioning*. It is characterized by simultaneous measurements made from a group of tracking stations, both known and unknown. The idea behind such measurements is that the satellite positions can be determined from the known stations, and the unknown positions are then deduced from satellite positions determined in this manner—see FIG. 7. Let  $\tau_j$  denote the epoch of time corresponding to a single observation  $l_{ij}$ ,—be it the range, direction, or range difference—from ground station  $P_i$  to satellite position  $S(\tau_j) = S_j$ . Each observation produces one observation equation of the following general form:

$$l_{ij} = f(\bar{r}_i, \bar{r}_j), \quad (17.18)$$

which does not depend explicitly on time. Considering a network of points  $P_i$  with either known or unknown positions  $\bar{r}_i$ , an array of unknown satellite positions  $S_j$  with coordinates  $\bar{r}'_j$ , and simultaneous observations  $l$  the above model becomes

$$l = f(x, x'). \quad (17.19)$$

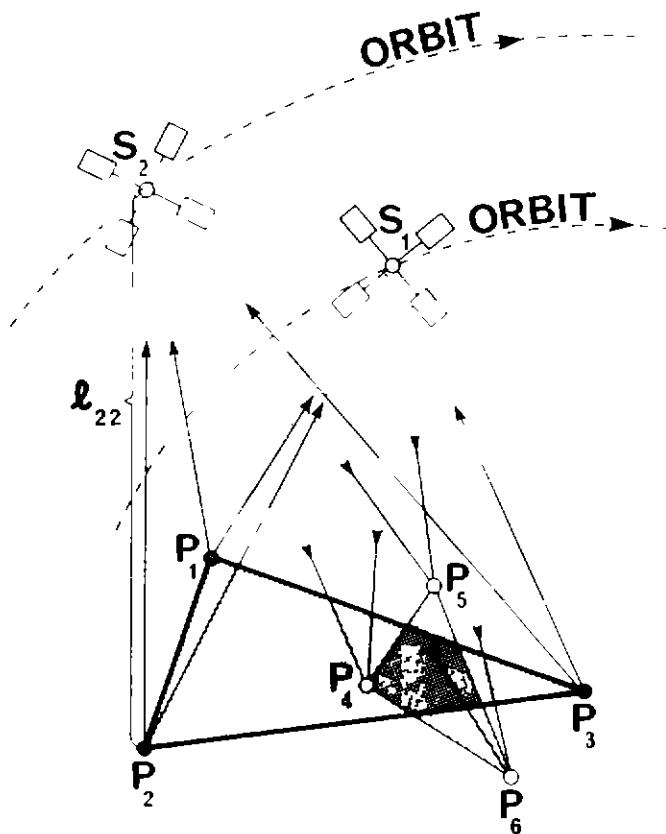


FIG. 17.7. Geometrical mode of satellite positioning. (● = known stations; ○ = unknown stations; ≈ = satellite position.)

Here  $\mathbf{x}$  denotes the vector of triplets  $(x, y, z)$ , and  $\mathbf{x}'$  the vector of triplets  $(x, y, z)'$ . The specific forms of the models  $f$  corresponding to different kinds of extraterrestrial measurements were introduced in §15.3 and §16.1 and will not be repeated here. Linearization of these inherently non-linear models  $f$  yields

$$\mathbf{r} = \mathbf{A}\delta\mathbf{x} + \mathbf{A}'\delta\mathbf{x}' + \mathbf{w}, \quad (17.20)$$

where  $\mathbf{A}$  and  $\mathbf{A}'$  are design matrices (see §12.1) referring to the tracking station positions ( $\mathbf{x}$ ) and satellite positions ( $\mathbf{x}'$ ) respectively. The  $\delta\mathbf{x}$ ,  $\delta\mathbf{x}'$  are the corrections to approximate values of the positions of both ground stations and satellites. Corrections  $\delta\mathbf{x}'$  are only nuisance parameters, usually much greater in number than  $\delta\mathbf{x}$  and, as such, are eliminated from the normal equations before solving for  $\delta\mathbf{x}$ , along the lines used in §12.2. The corrections  $\delta\mathbf{x}$  are clearly of two different kinds:  $\delta\mathbf{x}_1$ , the true unknowns that we seek, are corrections to the approximate values of the unknown tracking station positions, and  $\delta\mathbf{x}_2$ , corrections to the values of the known tracking stations. These have an a priori covariance matrix  $\mathbf{C}_{x_2}$  associated with them, and the solution is then sought using one of the techniques shown in §14.4.

Again, as mentioned in §15.3, critical configurations must be avoided. For instance, four tracking stations and a satellite position should not lie on the surface

of a sphere if directions to satellites are measured, or known and unknown tracking stations should not lie in the same plane. For details, see BLAHA [1971b].

(b) In the geometrical mode, the simultaneity of observations is of basic importance because there is no mathematical mechanism that links together observations made at two different instants. This synchronization of observations is sometimes difficult to achieve because of weather conditions and instrument limitations. To alleviate this (and other problems), the *kinematically constrained mode of simultaneous positioning* was devised. This mode makes use of the fact that the individual satellite positions  $S$ , along a given orbital arc are linked together by the physical laws governing the satellite motion (see §23.2).

To formulate the mathematical model for the kinematically constrained mode, one has to somehow encode the orbital information. It is convenient to do this in terms of six parameter state vectors  $z_k$  (cf. §10.3), containing three position components and three velocity components of the satellite. As will be shown in Chapter 23, the positions and velocities taken together define uniquely the motion of the satellite along at least a *short orbital arc*. A state vector  $z_k$  for the  $k$ th short arc can be formulated either in Cartesian coordinates, i.e.,

$$z_k = (x(\tau), y(\tau), z(\tau); \dot{x}(\tau), \dot{y}(\tau), \dot{z}(\tau))_k, \quad (17.21)$$

or, equivalently, in Keplerian orbital elements (cf. §15.3)

$$z_k = (a, e, i, \varpi, \varepsilon_\delta, \mu(\tau))_k. \quad (17.22)$$

The Keplerian elements (except  $\mu$ ) can be considered to remain practically constant for the duration of about one-eighth of an orbital revolution of the satellite [BROWN, 1970], and this is usually the meaning given to the above term of ‘short arc’. Thus, during one ‘short arc’, the mean anomaly  $\mu$  is the only parameter that varies with time. Even this variation is only linear (cf. §15.3).

Considering several short arcs, with numerous known satellite positions on them, one has several sextuples of orbital elements, and the vector of all the satellite positions  $x'$  can be written as a function of those sextuples

$$x' = g(z_1, z_2, \dots, z_n) = g(z). \quad (17.23)$$

After linearization we have, denoting  $\partial g / \partial z$  by  $B$ ,

$$r_g + \delta x' = B \delta z, \quad (17.24)$$

where  $r_g$  is the vector of model errors arising from the fact that even during a short orbital arc the Keplerian elements change little with time. If the model errors  $r_g$  are small relative to the observational errors, then the nature of the constraint stands out more explicitly. This setup has been investigated by SCHWARZ [1969].

Properly weighted with the inverse of  $C_g = E[r_g^T r_g]$ , eqn. (24) represents the kinematical constraint for the model (20) we have been looking for. Application of techniques for handling constrained models (see §14.5) and elimination of  $\delta x'$  yield

$$r = A \delta x + A'(B \delta z - r_g) + w. \quad (17.25)$$

Denoting  $\mathbf{r} + \mathbf{A}'\mathbf{r}_g$  by  $\tilde{\mathbf{r}}$  and  $\mathbf{A}'\mathbf{B}$  by  $\tilde{\mathbf{A}}$ , we get finally

$$\tilde{\mathbf{r}} = \mathbf{A} \delta \mathbf{x} + \tilde{\mathbf{A}} \delta z + \mathbf{w}. \quad (17.26)$$

One could, of course, go one step further and even eliminate  $\delta z$  so that only  $\delta \mathbf{x}$  remains. The treatment of  $\delta \mathbf{x}$  for both the known and unknown points is identical with that already mentioned in the context of the geometrical mode.

This version of kinematically constrained positioning became known as the *short-arc mode of satellite positioning*. The ‘short arc’ may be associated with a satellite pass over the tracking stations, in which case the situation portrayed in FIG. 8 occurs. It is interesting to realize that if one is also willing to include at least some time variations of the Keplerian elements (cf. §23.2) in the state vectors  $z_k$ , then the above described technique can be extended to *long (orbital) arcs* as well. These ‘long arcs’ may contain many orbital revolutions. We also note that the kinematically constrained mode of positioning is closely related to the ‘semidynamic mode’ of translocation described in §16.1. While in the former mode the satellite positions on one arc are constrained through the common Keplerian elements, the latter mode constrains the position through the biases common to each arc.

The main difference between the geometrical and kinematically constrained modes is that in the geometrical mode the coordinates of each satellite position are solved for independently, whereas in the kinematically constrained mode the satellite positions within the same orbital arc are linked together analytically. In spite of the obvious advantages of the latter mode, i.e., that the inclusion of the orbital information can only strengthen the solution, and that observations do not have to

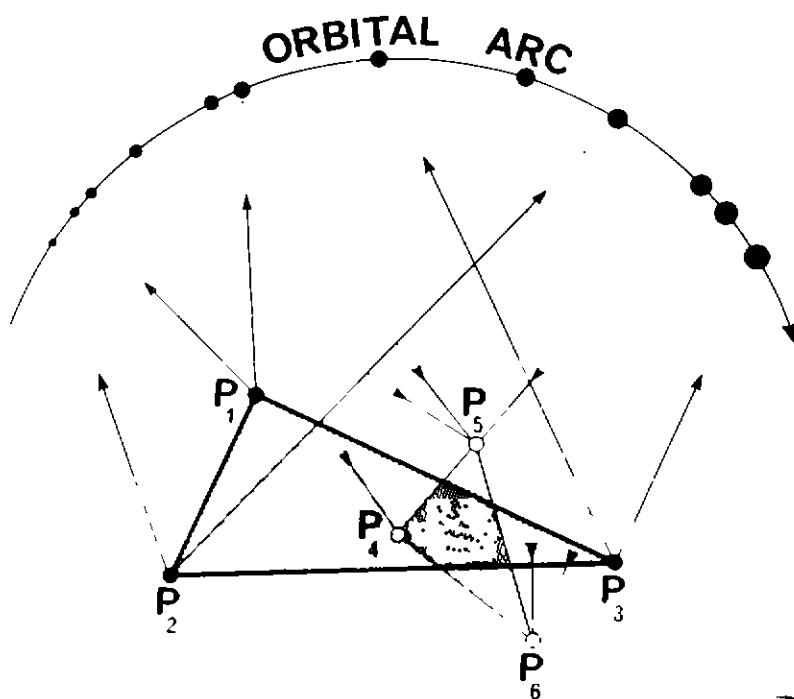


FIG. 17.8. Kinematically constrained mode of satellite positioning. (● = known stations; ○ = unknown stations.)

be made simultaneously, the geometrical mode has been used more extensively. This is especially the case in conjunction with direction measurements to balloon satellites which, due to their size, shape, and weight, have highly perturbed orbits [VEIS, 1960]. MUELLER [1974] reports on several satellite positioning systems that have used the geometrical mode.

The design matrices for the various aforementioned models are given in detail for various measurements to satellites in, e.g., MUELLER [1964] and KAULA [1966]. The interested reader is referred to BROWN AND TROTTER [1969] for an example of a TRANSIT satellite network determined in the short arc mode, where an accuracy of  $\sigma_x = \sigma_y = \sigma_z = 0.25$  m is reported.

The coordinate system in which the three-dimensional networks are positioned is imposed through the coordinates of the known points. This obviously can be done in the case when positions of some points are already known. But how is the whole process started? To begin with, we have to start with point positioning using the techniques of §15.2 and §15.3, then establish a few more points using the techniques of §16.1. The role of coordinate systems in these techniques has already been discussed.

Satellites can also be used in a dynamic mode. This mode has, however, a completely different objective from the geometrical and kinematically constrained modes in that, instead of seeking the best coordinates for tracking stations, it seeks to improve the knowledge of the gravity field, tidal phenomena, air density distribution etc. from observed perturbations of orbits. It uses long orbital arcs and it is through this mode that one can determine the CT coordinates of tracking stations; we shall briefly come back to this topic in §23.4.

The last topic that has to be discussed in the context of three-dimensional networks based on extraterrestrial observations is how to build up a network using pairs of stations positioned relatively by radio interferometry, satellite translocation, or inertial methods (see §16.1). If we assume the coordinate differences to be measured in redundant combinations, it is possible to construct a homogeneous network of points. The model,

$$\Delta \vec{r}_{ij} = \vec{r}_j - \vec{r}_i, \quad (17.27)$$

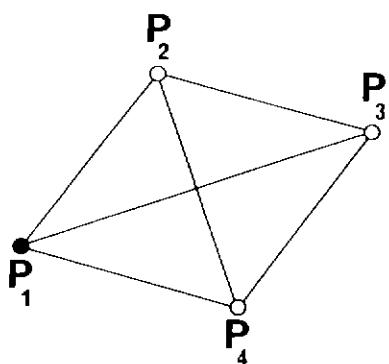


FIG. 17.9. Three-dimensional network composed of pairs of points with observed coordinate differences. (● = known station; ○ = unknown stations.)

relates the interstation vector  $\Delta\vec{r}_{ij}$  of observed coordinate differences, and  $\vec{r}_i$  and  $\vec{r}_j$  are the position vectors composed of the coordinates to be determined. For example, the network with one fixed point shown in FIG. 9 has  $6 \times 3 = 18$  observations and  $3 \times 3 = 9$  unknowns, resulting in a redundancy of 9. Given the covariance matrix for the observed coordinate differences, one can employ the methods given in Chapter 12 to solve directly for the coordinates.

#### 17.4. Assessment and merger of three-dimensional networks

The covariance matrix  $C_{\hat{x}}$  of the parameters (coordinates)  $\hat{x}$  contains all the information about the accuracy of the positions described by these coordinates  $\hat{x}$ . In the case of relative positioning (§16.1), we were able to interpret the covariance matrix as the confidence ellipsoid for the point that was being determined. The situation here is more complicated: we seek information about the position accuracy for a multitude of points in the network.

How can we obtain this information? Clearly, the source must, once again, be the covariance matrix  $C_{\hat{x}}$ . Within this matrix we can find—if necessary, after rearranging its rows and columns—a 3 by 3 submatrix  $C_{P_i}$  that belongs to the point  $P_i$ . This submatrix then plays the same role as the  $C_{\hat{x}}$  matrix in (16.21). In this case, what interpretation should be placed on the confidence ellipsoid described by (16.21)? While the answer to this question was quite clear in the case of relative positioning of one point with respect to another, here it is not.

To clarify the present situation, let us take the case of a terrestrial network first: when talking about the position of a point  $P_i$ , what is it taken with respect to?

(a) If the coordinate system is explicitly positioned and oriented at the origin  $P_0$  of the network, then all the points can be regarded as being positioned with respect to  $P_0$ . The initial point  $P_0$  then plays the role of the fixed point in the sense of relative positioning, and the confidence ellipsoids depict the uncertainty of the relative position of the points with respect to the origin. In spite of their relative character, they have become known as (out-of-context) absolute confidence ellipsoids of the network points. The general tendency of these absolute confidence ellipsoids is to grow in size with distance from  $P_0$ , and the ellipsoid associated with  $P_0$  degenerates to a point. This will be illustrated in FIG. 18.2 in the case of error ellipses.

(b) If the coordinate system has been positioned and oriented in another way (see §17.1), the confidence ellipsoids behave differently. When, for instance, the network has been adjusted through the use of inner constraints—instead of holding some parameters fixed in the adjustment—the confidence depicted by the ellipsoids refers to positions taken with respect to the centroid of the network. In this case, the trace of  $C_{\hat{x}}$  is also the minimum (cf. §14.5). Absolute confidence ellipsoids for the Canadian TRANSIT satellite network are shown in FIG. 10.

If one is interested in the accuracy of a relative position of any two points  $P_i, P_j$  within the network, one has to use the covariance law to derive it. The differences of

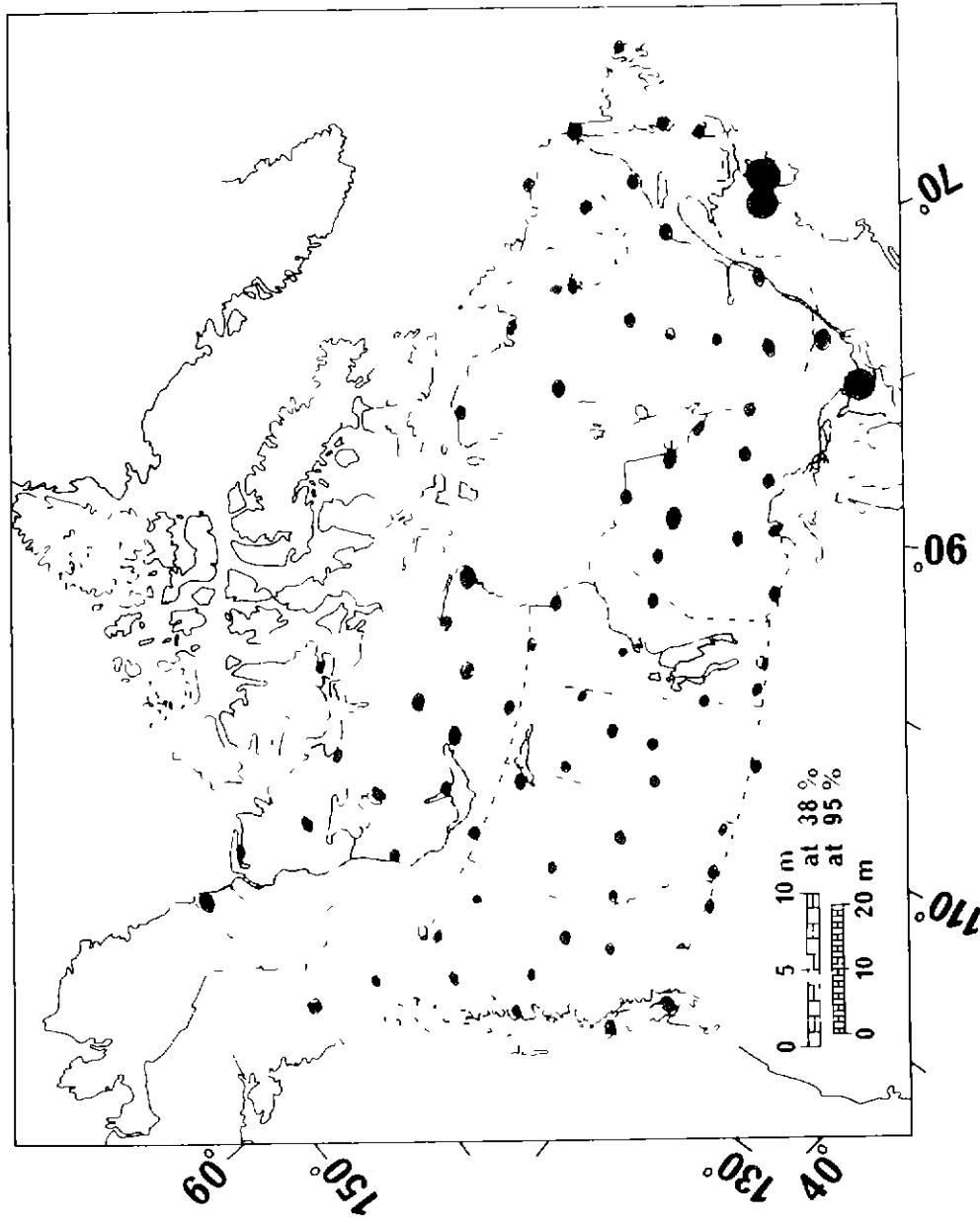


FIG. 17.10. Error ellipsoids of TRANSIT satellite determined points. (Courtesy of Geodetic Survey of Canada, DEPARTMENT OF ENERGY, MINES AND RESOURCES [1979], Ottawa, Canada.)

the adjusted coordinates of  $P_i$  and  $P_j$  can be written as

$$\Delta \hat{r}_{ij} = \begin{bmatrix} \Delta \hat{x}_{ij} \\ \Delta \hat{y}_{ij} \\ \Delta \hat{z}_{ij} \end{bmatrix} = \begin{bmatrix} \hat{x}_j - \hat{x}_i \\ \hat{y}_j - \hat{y}_i \\ \hat{z}_j - \hat{z}_i \end{bmatrix} = \mathbf{G} \begin{bmatrix} \hat{r}_j \\ \hat{r}_i \end{bmatrix}, \quad (17.28)$$

where  $\mathbf{G} = [\mathbf{I}_3^T - \mathbf{I}]$ , and  $\hat{r}_k = [\hat{x}_k, \hat{y}_k, \hat{z}_k]^T$  for  $k = i, \dots, j$ . Application of the covariance law yields

$$\mathbf{C}_{\Delta \hat{r}_{ij}} = \mathbf{G} \mathbf{C}_{\hat{P}_j \hat{P}_i} \mathbf{G}^T, \quad (17.29)$$

where  $\mathbf{C}_{\hat{P}_j \hat{P}_i}$  is a 6 by 6 covariance matrix:

$$\mathbf{C}_{\hat{P}_j \hat{P}_i} = \begin{bmatrix} \mathbf{C}_{\hat{P}_j} & \mathbf{C}_{ij} \\ \mathbf{C}_{ji} & \mathbf{C}_{\hat{P}_i} \end{bmatrix}. \quad (17.30)$$

The hypermatrix  $\mathbf{C}_{\hat{P}_j \hat{P}_i}$  is composed of two 3 by 3 covariance matrices for points  $P_j$  and  $P_i$ , selected from  $\mathbf{C}_{\hat{x}}$  the way described above, and two cross-covariance matrices  $\mathbf{C}_{ij}$ ,  $\mathbf{C}_{ji}$  also selected from the complete  $\mathbf{C}_{\hat{x}}$ . Evaluation of (29) yields

$$\boxed{\mathbf{C}_{\Delta \hat{r}_{ij}} = \mathbf{C}_{\hat{P}_j} + \mathbf{C}_{\hat{P}_i} - \mathbf{C}_{ij} - \mathbf{C}_{ji}.} \quad (17.31)$$

This 3 by 3 covariance matrix can once more be interpreted the same way as  $\mathbf{C}_{\hat{x}}$  was in §16.1. The corresponding confidence ellipsoid we thus obtain is the relative confidence ellipsoid of  $P_j$  with respect to  $P_i$ , or vice versa. The role of the expansion factor  $C_\alpha$  as a measure of probability is identical with that in §16.1.

To take proper statistical account of the existence of all points in the network when studying the confidence ellipsoids, either relative or absolute, requires that we employ the in-context approach developed in §13.4. The resultant expansion factor for the  $1 - \alpha$  in-context, or *simultaneous, confidence ellipsoid* is (cf. (13.41))

$$C_\alpha = (\xi_{\chi^2_3, 1-\alpha/N})^{1/2}, \quad (17.32)$$

if  $\sigma_0^2$  is used, or

$$C_\alpha = (3\xi_{F_{3,m-3N}, 1-\alpha/N})^{1/2}, \quad (17.33)$$

if  $\hat{\sigma}_0^2$  is used. Here,  $N$  is the number of unknown points in the network, and  $m$  is the number of observables used in the adjustment. The values of  $C_{0.05}$  for various cases would plot in a manner similar to that of FIG. 13.12. For a network of  $N = 50$  points and large  $m$ , the value of  $C_{0.05}$  goes to about 4.4; the size of a simultaneous confidence ellipsoid is then about 1.5 times that (2.8) of the confidence ellipsoid of one point taken out of context.

It should be remembered that, according to Bonferroni's inequality (13.21), the probability associated with simultaneous confidence ellipsoids is generally greater

than  $1 - \alpha$ , because of the neglect of the cross-covariances among the points in the network in the simultaneous probability statement. For example, if two independently determined sets of coordinates for the same points in the network do not simultaneously agree within the limits indicated by the confidence ellipsoids, then the explanation may lie in too stringent (i.e., too low) a significance level  $\alpha$  having been chosen in the first place.

In practice, we often have three-dimensional networks of different kinds covering the same area. We have seen one such case in §17.2, where we had a controlling terrestrial network and the densifying photogrammetrical network. It is, of course, highly desirable to exploit the strength of each kind of network (e.g., the local consistency of a terrestrial network versus the regionally homogeneous accuracy of a network based on extraterrestrial observations) by merging them together. Since the positions (coordinates) of network points usually come from different sources, the individual networks may refer to different coordinate systems. For a *merger of three-dimensional networks*, one generally has to go through various transformations such as we have seen in §17.2.

When the transformation, including its parameters, is known, the merging is a relatively simple task. To explain it properly, let us first agree, without any loss of generality, to designate the coordinate system of the first network as geocentric (N) and the other as geodetic (G)—see FIG. 11. Then the position  $\vec{r}_i^G$  of any point  $P_i$  of the second network can be transformed into the coordinate system of the first

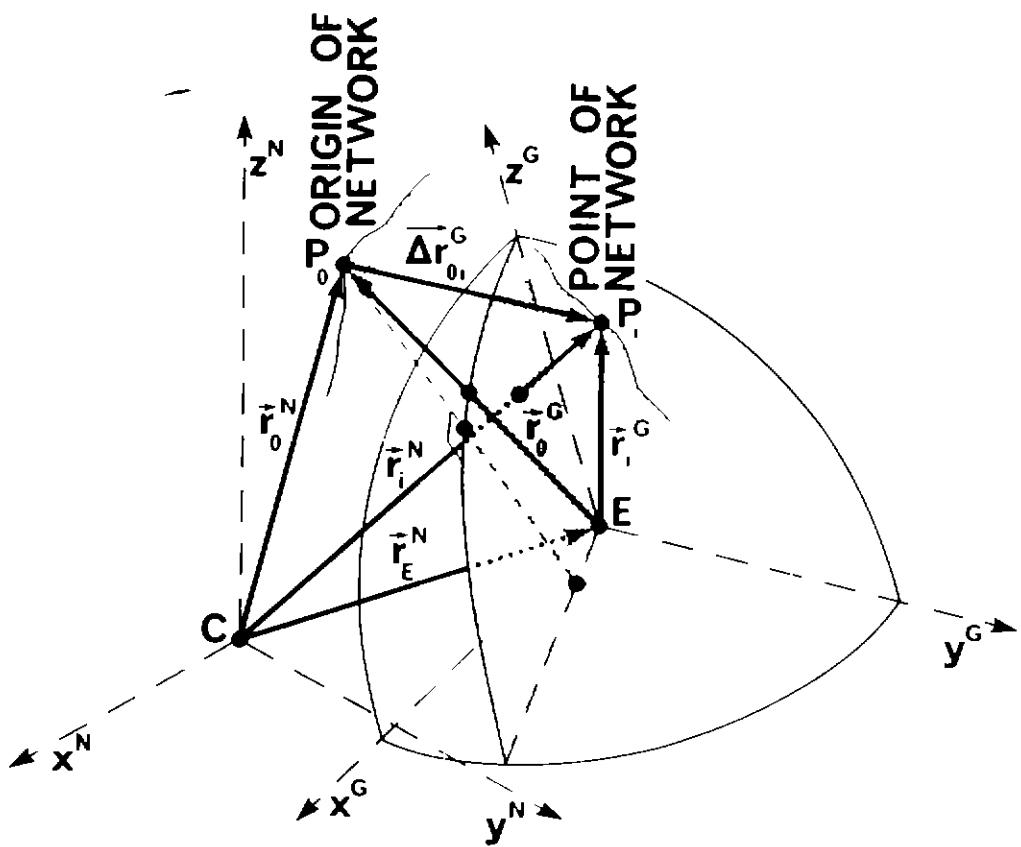


FIG. 17.11. Transformation of coordinates for merging of networks.

network by the following equation:

$$\bar{r}_i^N = \bar{r}_E^N + (1 + \kappa) R(\epsilon_x, \epsilon_y, \epsilon_z) \bar{r}_i^G, \quad (17.34)$$

where  $\epsilon_x, \epsilon_y, \epsilon_z$  are the misalignment angles of the two systems, and  $\kappa$  is the difference in the scales of the two networks. This is simply the known similarity transformation (17) written in a more explicit fashion.

Once the points of both networks are known in one coordinate system (N), the common points can be adjusted using proper weights. Each common point  $P_i$  gives six observation equations of the following kind:

$$\bar{r}_i^N + \hat{v}_i = \hat{r}_i^N, \quad \bar{r}'_i^N + \hat{v}'_i = \hat{r}'_i^N, \quad (17.35)$$

where  $\bar{r}'_i^N$  are the points belonging to the second network.

In practice, more often than not, the transformation parameters  $\bar{r}_E^N, \kappa, \epsilon_x, \epsilon_y, \epsilon_z$  in (34) are not known, or, even if they are, their accuracy may leave something to be desired. In such a case, one wants to estimate the best values of the parameters during the process of merging the networks. Clearly, if enough points are common to both networks, i.e., if there is a sufficient number of appropriately distributed points available whose coordinates are known in both systems, the parameters can be determined. Then the transformation itself becomes part of the model linking the quasi-observables, i.e., the coordinates of common points, with the unknown transformation parameters.

In practical applications, different models can be used according to what role the origin  $P_0$  of the second network plays in the transformation. BURŠA [1962] advocates the use of model (34) which implies rotations about the axes of the G system at  $E$ . MOLODENSKIJ ET AL. [1960] prefer to rotate the network around axes parallel to those of the G system at  $P_0$ . Their model reads (cf. FIG. 11)

$$\bar{r}_i^N = \bar{r}_0^N + (1 + \kappa) R(\epsilon_x, \epsilon_y, \epsilon_z) \Delta \bar{r}_{0i}^G. \quad (17.36)$$

VEIS [1960] uses a model identical with Molodenskij's, with the exception that the three rotations are carried out around the axes of the LG system at  $P_0$ .

The above models are non-linear in parameters and of the implicit variety; the solution of such models was discussed in Chapter 12. All three models give identical results for the rotations and scale differences. MUELLER AND KUMAR [1975] point out that the Burša model is particularly suitable for dealing with two networks with global coverage. In the situation of non-global extent (e.g., a national terrestrial network), it may be wiser to adopt either Molodenskij's or Veis's model, and seek rotation about the origin of the network.

Sometimes, it is possible to consider as known the mutual orientation of the two coordinate systems. This is what approximately happens, for instance, if the N system is the CT system and if the G coordinate system of the second network has been aligned to the CT system using the technique described in §15.4(c). If this is the case, then there is no need for introducing the rotations in the model. However,

rotational distortions of the second network with respect to the first, due to systematic errors in either network, may and often do exist. The average values of such distortions can then be modelled using the same models as before, except for a different interpretation of the rotation angles. This has been done for various combinations of networks, and the interested reader is referred to the works of, e.g., ANDERLE [1974], SCHMID [1974], MUELLER [1974], THOMSON AND KRAKIWSKY [1976], and HOTHÉM ET AL. [1978].

Clearly, if the systematic errors are present in one of the networks, and the orientation of the two systems cannot be considered known, more than the seven parameters are needed in the model. Specifically, to avoid network errors from getting aliased as the coordinate system rotations,  $\epsilon_x, \epsilon_y, \epsilon_z$ , a second set of rotations needs be included to model these distortions. The nature of these distortions is dictated by the manner in which the network has been established: networks established by three-dimensional terrestrial methods (§17.1) would, in general, have quite different systematic errors than those created by combining horizontal and vertical networks. Under these circumstances, HOTINE [1969] proposed to use the following model:

$$\begin{aligned} \bar{r}_i^N = & \bar{r}_E^N + \mathbf{R}_1(\epsilon_x) \mathbf{R}_2(\epsilon_y) \mathbf{R}_3(\epsilon_z) \\ & \times [\bar{r}_0^G + (1 + \kappa) \mathbf{R}_3(\pi - \lambda_0) \mathbf{R}_2(\frac{1}{2}\pi - \phi_0) \mathbf{P}_2 \mathbf{R}_H \mathbf{P}_2 \mathbf{R}_2(\phi_0 - \frac{1}{2}\pi) \\ & \times \mathbf{R}_3(\lambda_0 - \pi) \Delta \bar{r}_{0i}^G], \end{aligned} \quad (17.37)$$

where

$$\mathbf{R}_H = \begin{bmatrix} 1 & -\delta\alpha & \cos\alpha_0, \delta Z \\ \delta\alpha & 1 & \sin\alpha_0, \delta Z \\ -\delta Z/\cos\alpha_0 & 0 & 1 \end{bmatrix}, \quad (17.38)$$

and  $\alpha_0$  is the azimuth from the origin to the arbitrary point  $P_i$ ,  $\delta\alpha$  and  $\delta Z$  are the distortions in the azimuth and a zenith distance at the origin of the second network. Hotine acknowledges that a special estimation technique is required to separate the above two sets of rotations, i.e.,  $\epsilon_x, \epsilon_y, \epsilon_z$ , from  $\delta\alpha, \delta Z$ . THOMSON [1976] has worked out the details of an estimation technique in which the network is separated into two zones: an inner zone around the origin which is used to estimate  $\epsilon_x, \epsilon_y, \epsilon_z$ , with the remainder of the network being used to estimate  $\delta\alpha, \delta Z$ .

The KRAKIWSKY AND THOMSON [1974] model is based on the same general concepts as Hotine's with the only difference being that  $\mathbf{R}_H$  is replaced by

$$\mathbf{R}_K = \mathbf{R}_1(\delta\psi) \mathbf{R}_2(\delta\mu) \mathbf{R}_3(\delta A) \doteq \begin{bmatrix} 1 & -\delta A & \delta\mu \\ \delta A & 1 & -\delta\psi \\ -\delta\mu & \delta\psi & 1 \end{bmatrix}. \quad (17.39)$$

Here, the usual six parameters for the discordant G system are included, and four parameters (one scale difference  $\kappa$  and three rotations  $\delta\psi, \delta\mu, \delta A$ ) model the systematic errors in the terrestrial (second) network. As with the Hotine model, a special estimation technique is required to separate the two sets of rotations.

## CHAPTER 18

# HORIZONTAL NETWORKS

Horizontal networks were described in §7.1 from a general point of view; in this chapter we treat them in more depth. The first section describes how a horizontal datum is established, a task involving the selection of the size and shape of the reference ellipsoid and the specification of its position with respect to the earth. The second section presents the mathematical models used in obtaining horizontal coordinates from observed distances, directions, and azimuths. Also treated in this context are the numerical problems arising from dealing with large horizontal networks. In the third section, network assessment, design, extension, and densification are discussed. In addition, the concepts behind merging horizontal networks of different provenience are shown. The last section is devoted to a brief treatment of marine positioning, i.e., the determination of a horizontal position of a stationary or moving object on the sea surface.

### 18.1. Horizontal datum

In §7.1, a horizontal datum was defined as the appropriately positioned geodetic reference ellipsoid on which the horizontal coordinates  $\phi, \lambda$  of points in the network are reckoned. In order to transform these coordinates into other coordinate systems, the position of the horizontal datum must be known; the general idea of positioning of a horizontal datum with respect to the earth was discussed in §15.4(b) and (c). In addition, topocentric positioning (by means of the origin  $P_0$  of the network) was shown in §17.1 in the context of three-dimensional networks; this technique has been the standard for horizontal networks in the past [BOMFORD, 1971]. Since this *standard technique for horizontal datum positioning* is still being used, let us have a closer look at it here.

The six topocentric parameters needed (see §15.4) can be written as  $\phi_0, \lambda_0, h_0, \xi_0, \eta_0, \alpha_0$ ; they play the following roles:

(a)  $\phi_0, \lambda_0, h_0$  are the three geodetic coordinates of the initial point  $T_0 \equiv P_0$ , where  $h_0$  is the sum of the orthometric height  $H_0^O$  and the (relative) geoidal height  $N_0$  referred to the datum to be positioned (see (7.3)). The first two parameters,  $\phi_0, \lambda_0$ , specify a particular normal to the reference ellipsoid. Since the LA system (see FIG. 15.3) is fixed relative to the gravity field of the earth, once  $\xi_0$  and  $\eta_0$  are defined (see

below), it follows that the normal to the reference ellipsoid (specified by  $\phi_0, \lambda_0$ ) is also fixed with respect to the earth. The choice of  $h_0$  then fixes the depth of the ellipsoidal surface below or above the initial point ( $\Phi_0, \Lambda_0, H_0^0$ ); at this stage, the ellipsoid is free only to rotate around the normal.

(b)  $\alpha_0$  is the geodetic azimuth of one geodetic line joining  $P_0$  with another point in the network. This azimuth should satisfy (15.83), where  $\Lambda_0$  and  $A_0$  refer to the projection of  $P_0$  onto the geoid. The choice of  $\alpha_0$  removes the last remaining degree of freedom of the reference ellipsoid.

(c)  $\xi_0, \eta_0$ , are the two (relative) geoidal deflection components at  $P_0$  referred to the reference ellipsoid.

These six parameters are equivalent to the six geocentric parameters, three translations  $x_E, y_E, z_E$  and three misalignments  $\epsilon_x, \epsilon_y, \epsilon_z$ , introduced in §15.4(b). They cannot be, however, directly obtained from one another. Only the misalignment angles can be calculated from  $\phi_0, \lambda_0, \alpha_0, \xi_0, \eta_0$  using eqn. (15.81) (or more directly from eqn. (15.82)) if  $\Phi_0, \Lambda_0, A_0$  are also known. The three translations are most accurately and directly obtained from the comparison of a geocentric and a geodetic geoid solution (see §24.2).

The advantage that the topocentric parameters have over the geocentric is that they have a direct relation to quantities measured on the surface of the earth. How are these six initial parameters obtained? From the point of view of fixing the position of the datum, it is clearly immaterial where they come from as long as one is willing to regard them as constant. In the past, they used to be selected so as to make the reference ellipsoid (of preselected size and shape, normally thought to represent the size and shape of the earth as well as possible) fit the geoid in the best possible way in the region of interest (cf. §7.3 and §24.2). This used to be done so one could neglect all the reductions of geodetic observations for the effects of  $N, \xi, \eta$  (see §16.2).

If, in addition to the above described selection of  $\alpha_0$ , the deflection components  $\xi_0, \eta_0$  are specified so that they satisfy (15.84) and (15.85), which is what has been universally done, then the process automatically ensures the parallelism of the ellipsoid's minor axis with the  $z$ -axis of the CT system, as we have seen in §15.4(c). Under these conditions, the Laplace equation (15.83) is valid at any point of the network and can be used to obtain the Laplace azimuths (see (16.25)) from the observed astronomical azimuths. These help in strengthening the network but do not, as some scholars have believed, ensure the parallelism of the geodetic coordinate system to the CT system. In fact, we see that parallelism is established without the need of a network of points.

The above discussion should make us realize the importance of a proper selection of  $\alpha_0$  to ensure the parallelism. VANÍČEK AND WELLS [1974] give reasons why, in practice, parallelism cannot be achieved exactly, and a small misrotation  $\Delta_0$  of the G system around the ellipsoidal normal ( $\phi_0, \lambda_0$ ) with respect to the CT system, should always be expected (see FIG. 1). For example, for the NAD 27, a value for  $\Delta_0$  of between  $-0.2''$  and  $-0.3''$  has been estimated by WELLS AND VANÍČEK [1975]. The relation between  $\Delta_0$  and the misalignment angles was shown in eqn. (15.82).

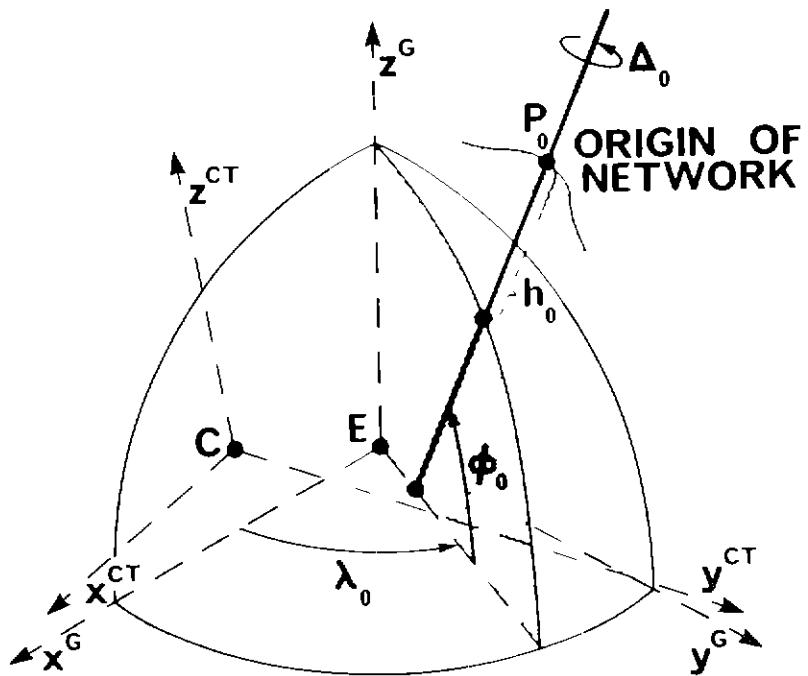


FIG. 18.1. Misalignment of a horizontal datum.

In the first approximation, it is not necessary to specify three of the six topocentric parameters: namely,  $N_0$ ,  $\xi_0$ ,  $\eta_0$ . This is possible because these quantities, at the origin as well as at other points, are not needed for the computation of horizontal coordinates  $\phi, \lambda$  from geodetic measurements in the initial stages of developing the network. In the first approximation, the effect of small geoidal heights and small deflections of the vertical on geodetic observations can be neglected. The remaining three parameters,  $\phi_0, \lambda_0, \alpha_0$ , however, have to be specified at the beginning to enable us to compute even the first approximations of the horizontal coordinates of the network points from distances and angles observed on the surface of the earth. In this case, orthometric instead of geodetic heights (above the reference ellipsoid) of points are used for the reduction of distances, and no corrections to horizontal angles are made. Also, when the geodetic work was confined to only one G system, no one had to worry about its relation to any other coordinate system, and it was not absolutely necessary to specify the accurate position of that system. Once other coordinate systems, such as a geocentric or another geodetic (belonging to another family), began to be used simultaneously, the proper transformation parameters had to be determined, and thus the  $\xi_0, \eta_0, N_0$  had to be specified.

Finally we may observe that the described classical method of positioning a datum does not rigorously fix the position of the G system with respect to the earth. This is because the LA system is fixed to the earth only up to the temporal variations of the earth's gravity field. Nevertheless, it is usual to regard the G system as fixed in the earth and the positions of control points in geodetic horizontal networks as expressed in this fixed geodetic coordinate system; then, whatever is done with the network does not affect the position of the coordinate system. The inevitable errors in the coordinate values originating from errors in observations as well as from

inaccurate computations and the neglect of various effects are then interpreted as simply errors in positions of these points. An example of such errors is given in FIG. 2. The addition of points or the readjustment of coordinates are also thought to have no effect on the coordinate system position unless some of the fundamental parameters are changed. Note that errors in astronomical coordinates influence only the geoid computations and enter into the above argument only as second-order errors in coordinates.

One alternative preferred by some researchers either explicitly, e.g., U.S. DEPARTMENT OF COMMERCE [1973] and JONES [1973], or implicitly, e.g., BURŠA [1965] and LAMBECK [1971], is the *floating datum*. In this option, the whole geodetic network, i.e., all the control points indiscriminately, are considered as defining the position of the geodetic datum and thus the geodetic coordinate system. The inherent problem with this approach is that, even though the geodetic coordinate system seems to be positioned with respect to the physical object, i.e., the network of points described by the adopted coordinate values of the points, this is not really the case. The positioning is really done through the geometrical representation of the physical object. Therefore, any errors in the initial determination of these coordinate values are transmitted into the position of the datum, and thus possible corrective measures

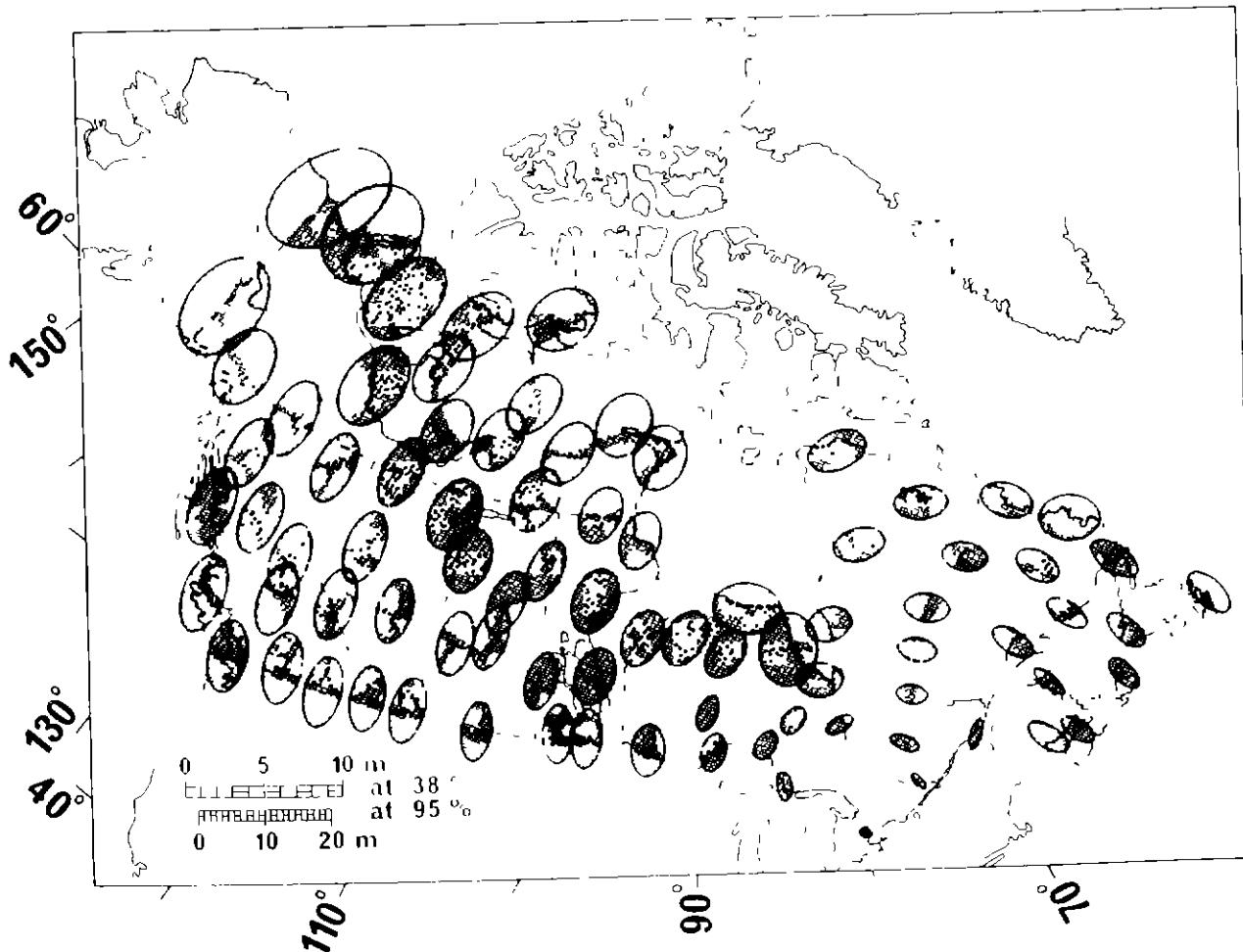


FIG. 18.2. Error ellipses corresponding to the adjustment of October, 1977. (Courtesy of Geodetic Survey of Canada, DEPARTMENT OF ENERGY, MINES AND RESOURCES [1979], Ottawa, Canada.)

bring about a change in the position of the coordinate system with respect to the earth. In addition, there is the rather unfortunate consequence of this definition in that the position of the geodetic datum with respect to the earth fluctuates (floats) with the addition of points to the network, with local readjustment, etc. This leaves us with a floating datum for which the transformation equations to another coordinate system are epoch dependent.

There is, however, something to be said in favour of the floating datum approach. In the standard technique, all the errors in positions, systematic as well as random, are associated with the network alone. In other words, only the geometrical representation of reality is considered distorted. On the other hand, in the floating datum approach, the errors are distributed evenly between the coordinate system and the network.

The compromise alternative, advocated by some geodesists, is *datum positioning by a set of selected points*, where a set of well-distributed control points is chosen and their coordinate values declared to define the position of the datum (with a preselected size and shape). The meaning of this definition is that the physical object consisting of the markers of the selected points is taken as the reference to which the coordinate system is then positioned. The spirit of this definition is identical with that used in defining the CIO (see §5.4). It is usually understood that in any subsequent computations, adjustments, or additions to the network, the coordinate values of these selected points will not be changed. This is the positioning technique used nowadays when satellite networks are merged with terrestrial networks (see §18.3); it is the natural technique to use, if the datum is to be positioned geocentrically.

## 18.2. Mathematical models and their solution

The observables in horizontal networks are horizontal angles  $\omega$  or directions  $d$ , spatial distances  $\Delta r$ , and astronomical azimuths  $A$ . The coordinates of the network points are usually sought, by means of the *adjustment of the network* (cf. §14.3), in the  $G$  system, i.e., as  $\phi$  and  $\lambda$  on the horizontal datum. Alternatively, the map coordinates  $(x, y)^M$  are sometimes sought instead.

If  $(\phi, \lambda)$  of the network points are the unknowns, then there are two ways to formulate the mathematical model relating these to the observables: the model is formulated either in three-dimensional space, or directly on the reference ellipsoid. In the former case, the observations are used in the model as they were observed on the surface of the earth (except for instrument and refraction corrections). The latter formulation uses observations reduced to the reference ellipsoid, as shown in §16.2. If a solution in the mapping plane is contemplated, the observations must first be reduced to the mapping plane (see §16.3).

In any of these three approaches, only approximate geodetic heights  $h$  of the network points are needed in the model. Conversely, none of these models yield any information about the vertical coordinates; this is the main difference from the three-dimensional networks treated in Chapter 17. It is useful to realize that, once

the horizontal coordinates and the approximate heights of the network points are known, corresponding horizontal angles, distances, and azimuths on the earth's surface can be obtained; this is the inverse to the problem shown in §16.2 and §16.3.

The *mathematical model for a horizontal network in three dimensions* is obtained directly from the model for three-dimensional networks (§17.1). Because accurate vertical angles  $\nu$  (or zenith distances  $Z$ ) and astronomical coordinates  $\Phi, \Lambda$  are not usually observed in horizontal networks, the observation equations pertaining to these quantities are not used. The remaining equations, i.e., those for the astronomical azimuth (17.2), direction (17.3), and spatial distance (17.6), however, contain coefficients which are functions of  $\nu$ . These coefficients must then be expressed as functions of other known quantities. The *astronomical azimuth observation equation* (for a horizontal network in three dimensions) is

$$r_{ij}^A = a_1 \delta\phi_i + a_2 \delta\lambda_i + a_4 \delta\phi_j + a_5 \delta\lambda_j + A_{ij}^{(0)} - A_{ij}, \quad (18.1)$$

where  $a_1, a_2, a_4, a_5$  are given by expressions found in TABLE 17.1, where the following replacements are made (for the derivation, see (15.4)):

$$\sin A \cos \nu = y^{\text{LA}} / \Delta r, \quad (18.2)$$

$$\cos A \cos \nu = x^{\text{LA}} / \Delta r. \quad (18.3)$$

The *direction observation equation* and *horizontal angle observation equation* are obtained in the same manner from corresponding expressions found in TABLE 17.1. The *spatial distance observation equation* reads as

$$r_{ij}^{\Delta r} = c_1 \delta\phi_i + c_2 \delta\lambda_i + c_4 \delta\phi_j + c_5 \delta\lambda_j + \Delta r_{ij}^{(0)} - \Delta r_{ij}. \quad (18.4)$$

Again,  $c_1, c_2, c_4, c_5$  are given in TABLE 17.1, where the functions of  $A, A'$ ,  $\nu$ , and  $\nu'$  are transformed using formulae (2) and (3). The coordinate differences required in the LA system are estimated from the approximate coordinates and predicted deflection components (see §24.3) where these approximate coordinates are obtained from repeated applications of relative positioning (as described in §16.2) to pairs of points in the network. This approach was tested at the U.S. National Geodetic Survey [VINCENTY AND BOWRING, 1978] within the context of the redefinition of the U.S. horizontal network. It is interesting to note that, in this approach, the gravity field parameters,  $N, \xi, \eta$ , are used in the evaluation of the coefficients, whereas in the other two approaches discussed below, these parameters are needed in the reduction of the observations to the reference ellipsoid.

The *mathematical model for a horizontal network on a reference ellipsoid* is based on the use of the geodesic curve between any two points on the ellipsoid. Thus azimuth, direction, and distance measurements must be reduced so that they refer to the geodesic; observation equations are then formulated on the ellipsoid. A special procedure must be used since, as we have seen in §16.2, it is difficult to formulate the equations in the explicit observable form. For example, it is difficult to express the ellipsoidal distance as a closed-form function of the coordinates  $\phi, \lambda$  of

the end points and then simply linearize it to obtain the observation equation in a differential form. The appropriate difference equations on the ellipsoid have to be formulated directly.

Conceptually, the ellipsoidal distance between  $P_i$  and  $P_j$  is written as (for simplicity, we omit the superscript E here)

$$S_{ij} = S(\phi_i, \lambda_i, \phi_j, \lambda_j). \quad (18.5)$$

Approximating the above with the linear part of a Taylor series results in

$$S_{ij} \doteq S(\phi_i^{(0)}, \lambda_i^{(0)}, \phi_j^{(0)}, \lambda_j^{(0)}) + \delta S = S^{(0)} + \delta S, \quad (18.6)$$

where  $S^{(0)}$  is the value of the ellipsoidal distance computed from the approximate coordinates  $\phi_i^{(0)}, \lambda_i^{(0)}$  and  $\phi_j^{(0)}, \lambda_j^{(0)}$ . The total difference  $\delta S$  is equal to

$$\begin{aligned} \delta S &= \frac{\partial S}{\partial \phi_i} \delta \phi_i + \frac{\partial S}{\partial \lambda_i} \delta \lambda_i + \frac{\partial S}{\partial \phi_j} \delta \phi_j + \frac{\partial S}{\partial \lambda_j} \delta \lambda_j, \\ &= c'_1 \delta \phi_i + c'_2 \delta \lambda_i + c'_4 \delta \phi_j + c'_5 \delta \lambda_j, \end{aligned} \quad (18.7)$$

where the coefficients are given in TABLE 1. The *ellipsoidal distance observation equation* (for a horizontal network on the ellipsoid) is then

$$r_{ij}^S = c'_1 \delta \phi_i + c'_2 \delta \lambda_i + c'_4 \delta \phi_j + c'_5 \delta \lambda_j + S_{ij}^{(0)} - S_{ij}, \quad (18.8)$$

where  $S_{ij}$  is the 'observed' value of the ellipsoidal distance, obtained by reducing (by means of (16.31)) the spatial distance  $\Delta r_{ij}$  measured on the earth's surface onto the ellipsoid.

The *geodetic azimuth observation equation* is developed in a similar fashion using the total differential of the azimuth. The result is (see, e.g., TOBEY [1928])

$$r_{ij}^\alpha = a'_1 \delta \phi_i + a'_2 \delta \lambda_i + a'_4 \delta \phi_j + a'_5 \delta \lambda_j + \alpha_{ij}^{(0)} - \alpha_{ij}, \quad (18.9)$$

where the coefficients are given in TABLE 1. The quantity  $\alpha_{ij}^{(0)}$  is the value of the azimuth computed from the approximate coordinates, and  $\alpha_{ij}$  is the 'observed' value obtained by reducing the astronomical azimuth  $A_{ij}$  to the ellipsoid through the use of (16.29).

TABLE 18.1  
Coefficients of the design matrix for horizontal networks on the reference ellipsoid,  
according to HELMERT [1880]. (Prime denotes quantities related to the second point,  $P_j$ )

Unknown	Subscript	'Observed' $\alpha$ or $d$ (or $\omega$ )		'Observed' $S$
		$a'$	$c'$	
$\phi_i$	1	$M \sin \alpha / S$	$-M \cos \alpha$	
$\lambda_i$	2	$N' \cos \alpha' \cos \phi' / S$	$N' \sin \alpha' \cos \phi'$	
$\phi_j$	4	$M' \sin \alpha' / S$	$-M' \cos \alpha'$	
$\lambda_j$	5	$-N' \cos \alpha' \cos \phi' / S$	$-N' \sin \alpha' \cos \phi'$	

The *direction observation equation* is simply

$$r_{ij}^d = a'_1 \delta\phi_i + a'_2 \delta\lambda_i + a'_4 \delta\phi_j + a'_5 \delta\lambda_j - \delta\Omega_i + \alpha_{ij}^{(0)} - d_{ij} - \Omega_i^{(0)}, \quad (18.10)$$

where the orientation unknown  $\Omega_i = \Omega_i^{(0)} + \delta\Omega_i$  (see FIG. 3) plays the same role as in (17.3). The *horizontal angle observation equation* is obtained from the above as

$$\begin{aligned} r_{ijk}^\omega = & (a'_1(k) - a'_1(j)) \delta\phi_i + (a'_2(k) - a'_2(j)) \delta\lambda_i + a'_4(k) \delta\phi_k \\ & + a'_5(k) \delta\lambda_k - a'_4(j) \delta\phi_j - a'_5(j) \delta\lambda_j + \omega_{ijk}^{(0)} - \omega_{ijk}, \end{aligned} \quad (18.11)$$

where  $\omega_{ijk}^{(0)}$  denotes the value of the angle on the ellipsoid computed from the approximate coordinates, and  $\omega_{ijk}$  denotes the observed value after reduction from the terrain onto the ellipsoid. The above family of observation equations for horizontal networks is used in computer programmes GALS [MCLELLAN ET AL., 1970] and the Land Registration and Information Service (LRIS) package [KNIGHT AND MEPHAM, 1978].

The last is the *mathematical model for a horizontal network on a conformal mapping plane*; it is the simplest of them all. Realizing that the chord distance  $l_{ij}$  (cf. §16.3) between points  $P_i$  and  $P_j$  is given simply as (see FIG. 4)

$$l_{ij} = [(x_j - x_i)^2 + (y_j - y_i)^2]^{1/2}, \quad (18.12)$$

after linearization one obtains the following *chord distance observation equation* (for a horizontal network on a conformal mapping plane):

$$r_{ij}^l = -\sin t_{ij} \delta x_i - \cos t_{ij} \delta y_i + \sin t_{ij} \delta x_j + \cos t_{ij} \delta y_j + l_{ij}^{(0)} - l_{ij}. \quad (18.13)$$

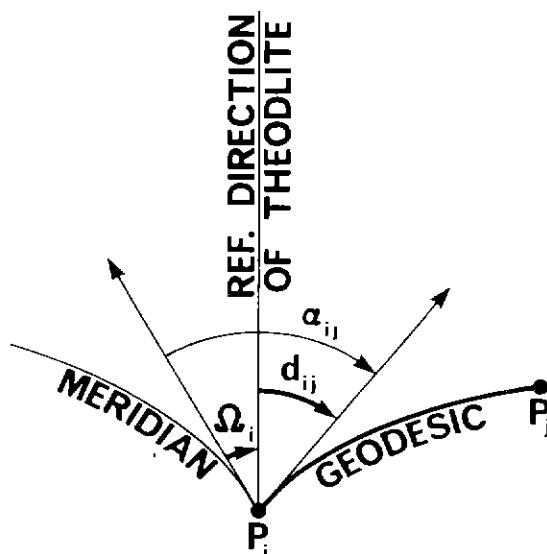


FIG. 18.3. Orientation of horizontal directions on the reference ellipsoid.

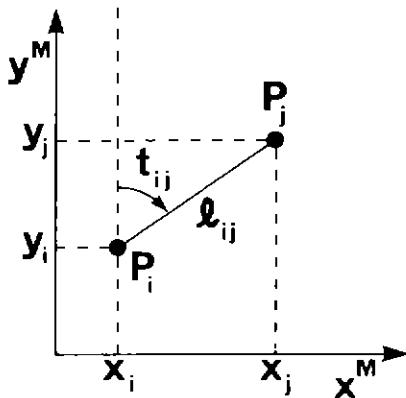


FIG. 18.4. Mathematical model on a mapping plane.

Here,  $t_{ij}$  is the grid azimuth of the chord;  $l_{ij}^{(0)}$  is the chord length computed from the approximate coordinates  $x_i^{(0)}, y_i^{(0)}$  and  $x_j^{(0)}, y_j^{(0)}$ ; and  $l_{ij}$  is the ‘measured’ value of the distance reduced from the terrain to the conformal mapping plane following the procedures outlined in §16.2 and §16.3.

The *grid azimuth observation equation* is

$$r'_{ij} = -\frac{\cos t_{ij}}{l_{ij}} \delta x_i + \frac{\sin t_{ij}}{l_{ij}} \delta y_i + \frac{\cos t_{ij}}{l_{ij}} \delta x_j - \frac{\sin t_{ij}}{l_{ij}} \delta y_j + t_{ij}^{(0)} - t_{ij}, \quad (18.14)$$

where  $t_{ij}$  is the ‘observed’ and  $t_{ij}^{(0)}$  the computed values of the grid azimuth. This equation is obtained as the linearized version of the well-known equation for the grid azimuths:

$$t_{ij} = \arctan[(x_j - x_i)/(y_j - y_i)]. \quad (18.15)$$

The *observation equations for a direction and a horizontal angle* then follow from (14) along the same lines as those in the previous mathematical model. This last model, for different conformal mappings, is the basis of several computer programmes such as GANET [BEATTIE, 1978] and TRAV10 [SCHWARZ, 1978].

One is now tempted to ask if any one of the three models is preferable. First of all, it is easy to show that the ellipsoidal and the conformal mapping plane models are equivalent (up to the effect of linearization), if the length elements  $dx$  and  $dy$  are equal to the length elements  $dS_\phi$  and  $dS_\lambda$  (see FIG. 16.17) on the ellipsoid. The first two models, i.e., the three-dimensional and the ellipsoidal, must be equivalent since they express the relation between the same observables and coordinates in the same system. The only difference is that the former model includes the gravity effects while the latter does not; in the latter model, the gravity field effects are accounted for in the reduction of the observations to the reference ellipsoid. Thus the results, i.e., the positions of the network points, obtained from the three models must be equivalent even though the third yields coordinates in a different coordinate system.

In practice, it is usual to use either seconds of arc (for the angular quantities) together with metres or centimetres (for the linear quantities), or radians together

with a relative (unitless) measure for distances. These units keep the magnitudes of quantities expressed in this way comparable and thus help in preventing the matrix of normal equations from becoming ill conditioned.

The above mathematical models may be expanded to include additional parameters parameterizing certain systematic effects in the network. One such parameterization was shown in §17.3, within the context of three-dimensional networks. In horizontal networks, systematic effects in need of parameterization are the effects of refraction on horizontal directions and spatial distances. As far as the latter is concerned, it has been known for some time that distances measured with different instruments or at different times are affected differently. This effect is handled simply by introducing one or more unknown parameters in each distance observation equation (see, e.g., ANGUS-LEPPAN [1972]). These parameters may then be treated as nuisance parameters and are eliminated, together with the orientation unknowns  $\Omega$ , before the solution for the coordinate increments is attempted; the elimination technique was shown in §14.5.

Some continents or countries have very extensive horizontal networks (cf. FIG. 7.2). Compared with existing three-dimensional networks, the number of points in a horizontal network may be very large indeed. For example, the Canadian Maritime Provinces' network comprises about 40000 points of mixed orders (FILA AND CHAMBERLAIN, 1978], and the U.S. national network has almost 250000 points of first order [ISNER AND YOUNG, 1978]. Since each point yields two unknowns—the coordinate increments  $\delta\phi, \delta\lambda$  (or  $\delta x, \delta y$ )—one could be faced with the necessity of solving for several hundred thousand unknowns, even if the nuisance parameters had already been eliminated. This is not a trivial task even for the most modern computers and a special strategy for a solution must be adopted.

Helmert [WOLF, 1978] was the first to outline this problem's solution, which has become known as *Helmert blocking*. It can be described in five steps.

(a) In the first step, only normal equations belonging to one block (e.g., block 1 in FIG. 5) are considered. These are partitioned into two parts: namely,

$$\begin{bmatrix} \mathbf{N}_x & | & \mathbf{N}_{xy} \\ \hline \mathbf{N}_{yx} & | & \mathbf{N}_y \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_x \\ \hline \boldsymbol{\delta}_y \end{bmatrix} + \begin{bmatrix} \mathbf{u}_x \\ \hline \mathbf{u}_y \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \hline \mathbf{0} \end{bmatrix}, \quad (18.16)$$

where the subscript  $x$  indicates the points within the block, while  $y$  indicates junction points, i.e., points common with adjacent blocks (see  $y_1$  and  $y_2$  in FIG. 5).

(b) Then the reduced normal equations for junction points  $y$  are formed by eliminating inner points  $x$  (see §3.1):

$$\bar{\mathbf{N}}_y \boldsymbol{\delta}_y + \bar{\mathbf{u}}_y = \mathbf{0}, \quad (18.17)$$

where

$$\bar{\mathbf{N}}_y = \mathbf{N}_y - \mathbf{N}_{yx} \mathbf{N}_x^{-1} \mathbf{N}_{xy}, \quad (18.18)$$

and

$$\bar{\mathbf{u}}_y = \mathbf{u}_y - \mathbf{N}_{yx} \mathbf{N}_x^{-1} \mathbf{u}_x. \quad (18.19)$$

Note that at this stage, before involving further blocks, a complete solution for the

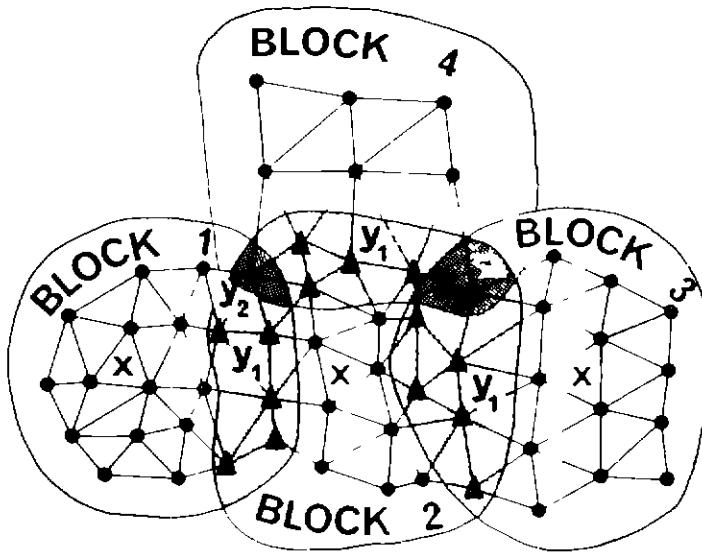


FIG. 18.5. Helmert blocking of a network. ( $\blacktriangle$  = junction points;  $\bullet$  = inner points.)

block is possible. The coordinates of junction points  $y$  are obtained from (17), and substitution into the first equation of (16) yields the solution for the inner unknowns  $\delta x$ .

(c) When dealing with the entire network, instead of only one block, we distinguish junction points of different levels, depending on how many blocks they belong to. Thus we have junction points of the first level  $y_1$ , of the second level  $y_2$  (see FIG. 5), and so on. Clearly, each block may participate in more than one level (e.g., block 2 in FIG. 5), so it is expedient to partition the system (16) even further. Considering the  $i$ th block with junction points of two levels, we have

$$\begin{bmatrix} \bar{\mathbf{N}}_{x_i} & \bar{\mathbf{N}}_{x_i, y_1} & \bar{\mathbf{N}}_{x_i, y_2} \\ \bar{\mathbf{N}}_{y_1 x_i} & \bar{\mathbf{N}}_{y_1} & \bar{\mathbf{N}}_{y_1, y_2} \\ \bar{\mathbf{N}}_{y_2 x_i} & \bar{\mathbf{N}}_{y_2 y_1} & \bar{\mathbf{N}}_{y_2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_i \\ \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{x_i} \\ \mathbf{u}_{y_1} \\ \mathbf{u}_{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \quad (18.20)$$

(d) As before, the first step in dealing with this new system is to eliminate  $\hat{\mathbf{x}}$ , and get the *main system of normal equations* for the first level junction unknowns. A similar reduced system can be obtained for the same junction points from the other pertinent blocks. It is at this level that other main systems of normal equations reduced in this manner, from other blocks involving the same junction points  $y_1$ , are merged. This is repeated for all the blocks that contain first-level junction points. Then the main system of normal equations is formed as

$$\tilde{\mathbf{N}}_{y_1} \hat{\mathbf{y}}_1 + \tilde{\mathbf{u}}_{y_1} = \mathbf{0}, \quad (18.21)$$

where

$$\tilde{\mathbf{N}}_{y_1} = \sum_i (\bar{\mathbf{N}}_{y_1})_i \quad \text{and} \quad \tilde{\mathbf{u}}_{y_1} = \sum_i (\bar{\mathbf{u}}_{y_1})_i, \quad (18.22)$$

which involves all the  $y_1$  junction points.

(e) Then  $y_1$  is eliminated to give the main system of normal equations for the second level junction points. Again, it is possible to add other reduced normal equations stemming from other blocks containing second level junction points. This procedure continues to the highest level of junction points. Substitution back into the unreduced systems of equations recovers, in reverse order, the junction points of lower levels and, eventually, even the inner points.

Matrices of normal equations for horizontal networks contain many zero elements. It is then possible to rearrange the sequence of points in the network so as to minimize the width of the band along the main diagonal of the matrix that contains non-zero elements, called the *profile of the matrix*, and make the computations more economical [SNAY, 1976]. Also, the reduction of the Helmert blocking matrix systems should be done by the use of Cholesky's square-root method to minimize the effect of round-off errors [MEISSL, 1978]. Computer programmes using Helmert's blocking strategy and other features to minimize computing cost have been successfully developed by, e.g., ISNER [1978]. KNIGHT AND MEPHAM [1978] worked out an alternative approach based on the summation of normal equations (see §14.6).

There is no reason why Helmert's blocking or the summation of normal equations approach could not be used for solving models for three-dimensional networks as well. So far, the need has not arisen, however, because the number of unknown point coordinates has been very much smaller than that of the horizontal network.

### 18.3. Assessment, expansion, and merger of horizontal networks

Even when the positions of network points have been determined, the network may still have to undergo various changes. It may get densified, expanded into new areas, and it may have to be merged with another network covering the same region. All these tasks are, of course, not unique to horizontal networks: they exist naturally in the realm of three-dimensional networks, as we have seen in §17.4, as well as in the realm of height networks. The ideas pertaining to densification and expansion were not introduced earlier, however, because the formalism for these has historically been developed and elaborated upon within the context of horizontal networks.

For any of these three tasks, it is essential to have an a priori assessment of the accuracy of the network. This assessment should be, and usually is, done along two parallel lines: assessment of the effect of random errors and assessment of systematic distortions. Let us begin with the assessment of random errors. As in the case of three-dimensional networks (cf. §17.4), these errors can be fully characterized by the covariance matrix  $C_{\hat{x}}$  of the estimated coordinates. This matrix, in turn, can be interpreted in terms of absolute or relative confidence ellipses, as shown in §16.2. Again, the absolute confidence ellipses have a tendency to grow as one goes away from the origin of the network; this tendency is rather nicely illustrated on the example of the Canadian first-order (terrestrial) horizontal network in FIG. 2.

It is because of this tendency to grow, which overwhelms all the other information contained in them, that absolute confidence ellipses are really of little use for assessing the random errors in networks. Of more value are relative confidence ellipses. These are obtained from the covariance matrix of the coordinates following

the approach shown in §17.4, except that here the number of dimensions is two. One possible geometrical interpretation of a relative confidence ellipse is shown in FIG. 6. It is interesting to see that the relative confidence ellipse also reflects very simply the standard deviations  $\sigma_i$  in the adjusted distance  $\hat{l}$  and  $\sigma_{\hat{a}}$  in the adjusted azimuth  $\hat{a}$ . Often it is the relative accuracy  $\sigma_i/l$  that determines the classification of the network [DEPARTMENT OF ENERGY, MINES AND RESOURCES, 1973].

As with the single point confidence ellipse, the relative confidence ellipse is not concerned with the other relative accuracies in the network. To take the simultaneous existence of the other points in the network into account, and to obtain the in-context, or *simultaneous, confidence ellipses*, be they absolute or relative, requires that we employ the in-context approach developed in §13.3. The only things that get changed are the expansion factors  $C$  ((13.36) and (13.37)) that read (note that  $\alpha$  here denotes the probability level and not an azimuth)

$$C_\alpha = \left( \xi_{\chi^2_2, 1-\alpha/N} \right)^{1/2}, \quad (18.23)$$

if  $\sigma_0^2$  is used, or

$$C_\alpha = \left( 2 \xi_{F_{2,m-2N}, 1-\alpha/N} \right)^{1/2}, \quad (18.24)$$

if  $\delta_0^2$  is used. Here,  $N$  is the number of points (or pairs of points in the case of relative ellipses) that are examined together, and  $m$  is the number of observables used.

Expansion factors  $C_{0.05}$  for various  $N$  and  $m$  plot in a manner similar to FIG. 13.12. As an example,  $C_{0.05}$  for  $N=50$  and large  $m$  tends toward about 3.8, a value more than 1.5 times larger than that (2.45—cf. §16.2) for the *out-of-context ellipse*. The sizes of the different absolute error ellipses are given in FIG. 7. Finally, recall that Bonferroni's inequality (13.21) shows that the probability associated with the simultaneous ellipses is even larger than  $1 - \alpha/N$ .

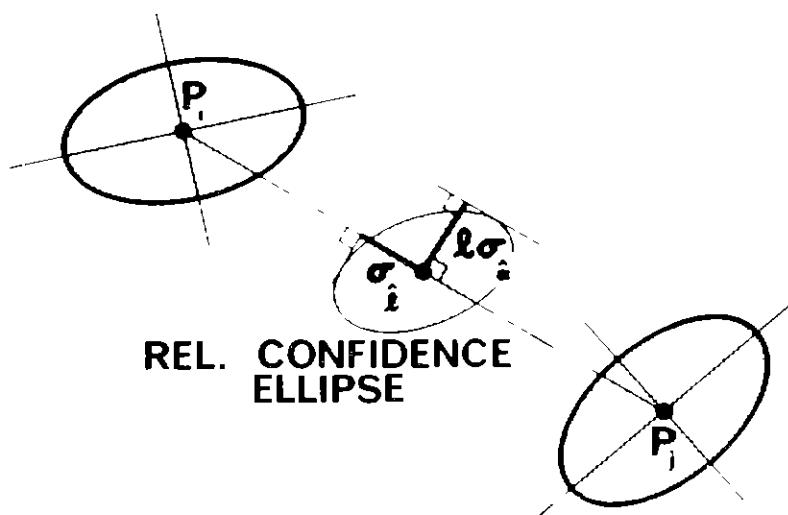


FIG. 18.6. Absolute and relative confidence ellipses.

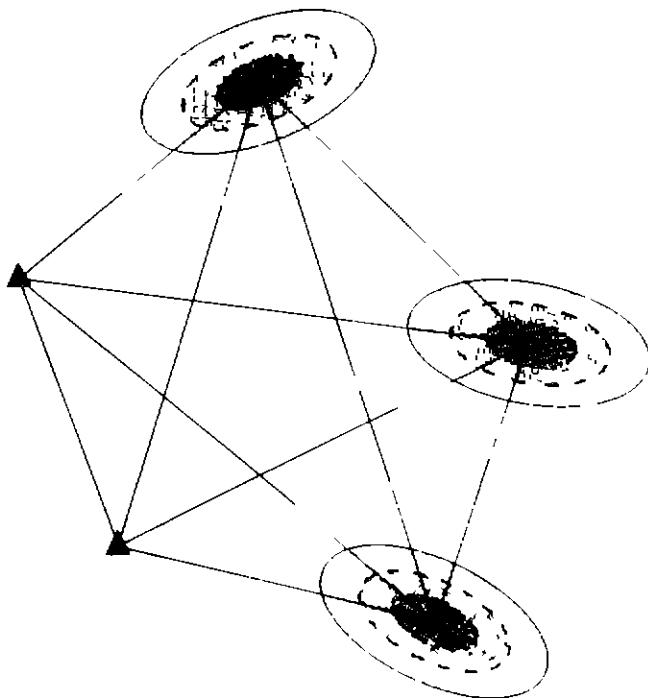


FIG. 18.7. Absolute confidence ellipses. (...standard ellipses, -·- 0.95 out-of-context ellipses; — 0.95 simultaneous ellipses.)

FIG. 8 shows the random error propagation patterns in horizontal networks of different shapes and kinds of observables. The simulated shapes are evident from the figure and so should be the kinds of observables: distances, directions, and azimuths (denoted by arrowheads). Points, whose coordinates have been assumed known within the standard confidence circle of radius of 0.5 m, are denoted by black triangles. When two such control points are used, a 95% relative confidence circle of radius 22 cm is assumed. The standard deviation of the observed directions and azimuths is assumed to be two seconds of arc corresponding to about  $10^{-5}$  radians; the relative accuracy of distances is assumed to be  $10^{-5}$ . Plotted for each network, in a continuous diagrammatic manner, are the values of major semi-axes of the 95% out-of-context point confidence ellipses. Also plotted are the 95% out-of-context relative confidence ellipses for pairs of adjacent points.

One can clearly see that, from the point of view of the *strength of the network*, as understood by ASHKENAZI AND CROSS [1972], the areal configuration is the best: the random errors propagate very slowly and rather uniformly. Generally, a network controlled at the edges through known control points is preferable to a network anchored at only one point. This pattern is made even clearer when the point position errors are recapitulated in FIG. 9. The rate of increase of the absolute confidence ellipses for a real network should fall between the two bottom curves in FIG. 9(a). For further examples, the interested reader is referred to CHRZANOWSKI AND KONECNY [1965].

Let us make a small detour here and have a look at the *optimum design analysis of horizontal networks*. As discussed in §14.1, it is possible to forecast the covariance

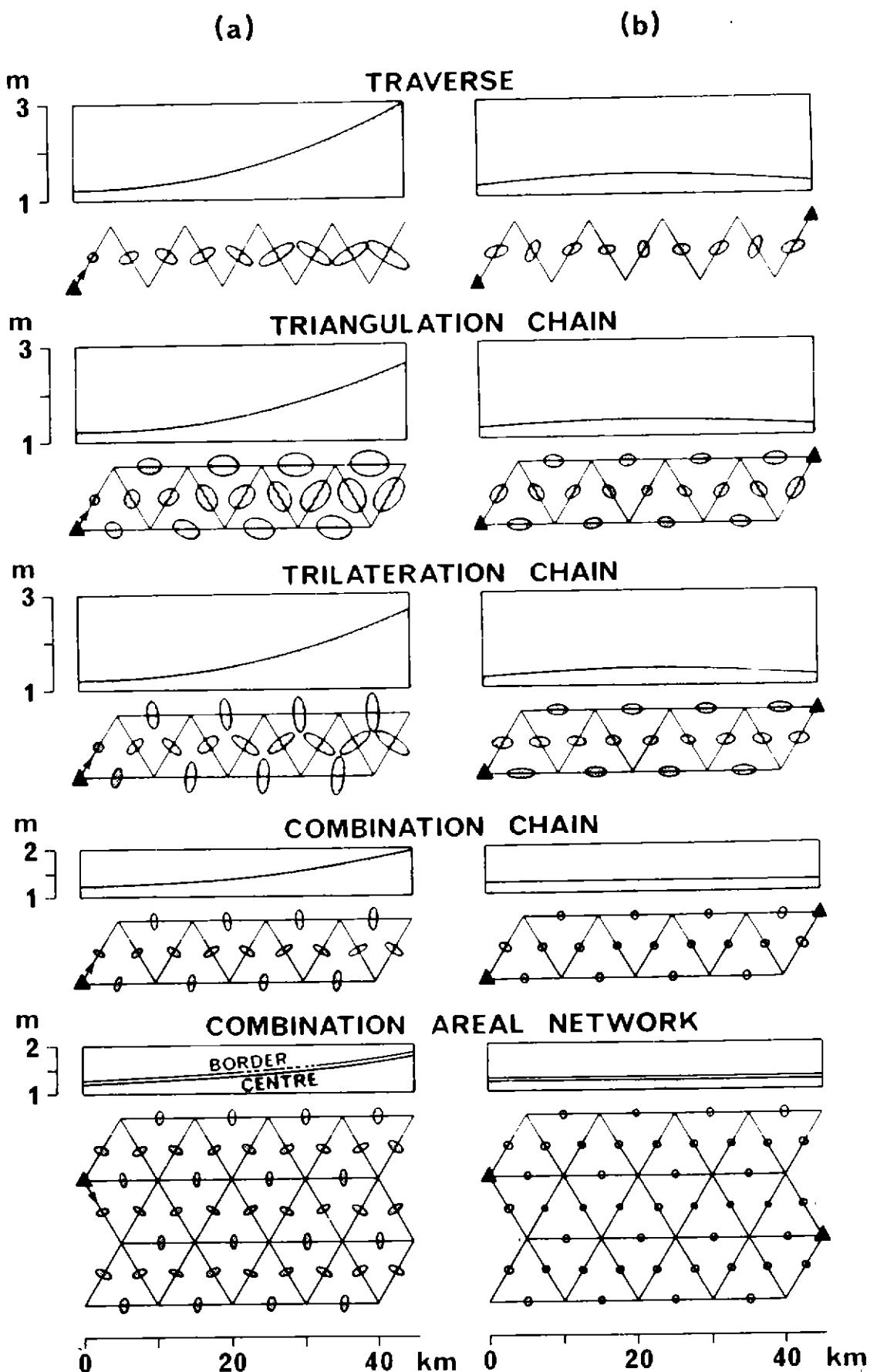


FIG. 18.8. Error propagation in horizontal networks. (For an explanation, see the text.)

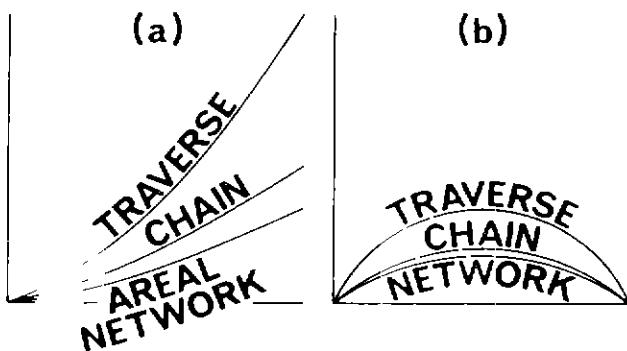


FIG. 18.9. Position error propagation.

matrix  $\mathbf{C}_x$  of the unknown horizontal coordinates before the measurements are actually made. This is accomplished simply by specifying both the network configuration (through the design matrix  $\mathbf{A}$ ) and the proposed accuracy of the observations (through the matrix  $\mathbf{C}_l$ ). Inverse procedures, whereby given  $\mathbf{C}_x$ ,  $\mathbf{A}$  and  $\mathbf{C}_l$  are sought, exist. Their drawback is that they employ a trial and error approach requiring time consuming repetitions (trials). To speed up this procedure, interactive computer graphics (see §14.1) have been employed where a cathode-ray tube linked to a computer is used to display the network configuration, along with confidence ellipses for the points, in real time for each trial.

To end the discussion of errors, we should mention at least two other concepts related to the characterization of the accuracy of horizontal networks by confidence ellipses. As we have seen earlier in this section, the absolute ellipses grow in size away from the initial point held fixed in the adjustment and thus depend on the choice of origin. Is there a way of avoiding this unfortunate property, other than relying on relative ellipses?

One way is to leave the origin (regarded as errorless) unspecified in the model for the network. The resultant rank deficient network is then adjusted using inner constraints (see §14.5). The specific form of inner constraint design matrix  $\mathbf{D}_i$  had the form of (17.14) with the only difference being the dimensionality. This gives  $\min_{\hat{\mathbf{x}}} \text{tr} \mathbf{C}_{\hat{\mathbf{x}}}$  (14.114), and the absolute confidence ellipses are all referred to the centroid of the points in the network. The relative accuracy, though, remains unchanged; clearly, the inner constraints, or the pseudoinverse, do not solve the problem of the dependency of absolute ellipses on the choice of the initial point but only shift the initial point to coincide with the centroid.

BAARDA [1973] approached the problem from a different angle: he formulated *S-transformations*, which are matrices that transform the confidence ellipses from one shape and size to another with the shift of the initial point or with the replacement of one or more initial points by other initial points. The latter transformation, of course, changes even the relative error ellipses in the network (see FIG. 10).

We have now come to the point where it makes sense to explore other methods for analysing the strength of horizontal networks. These methods are equally applicable

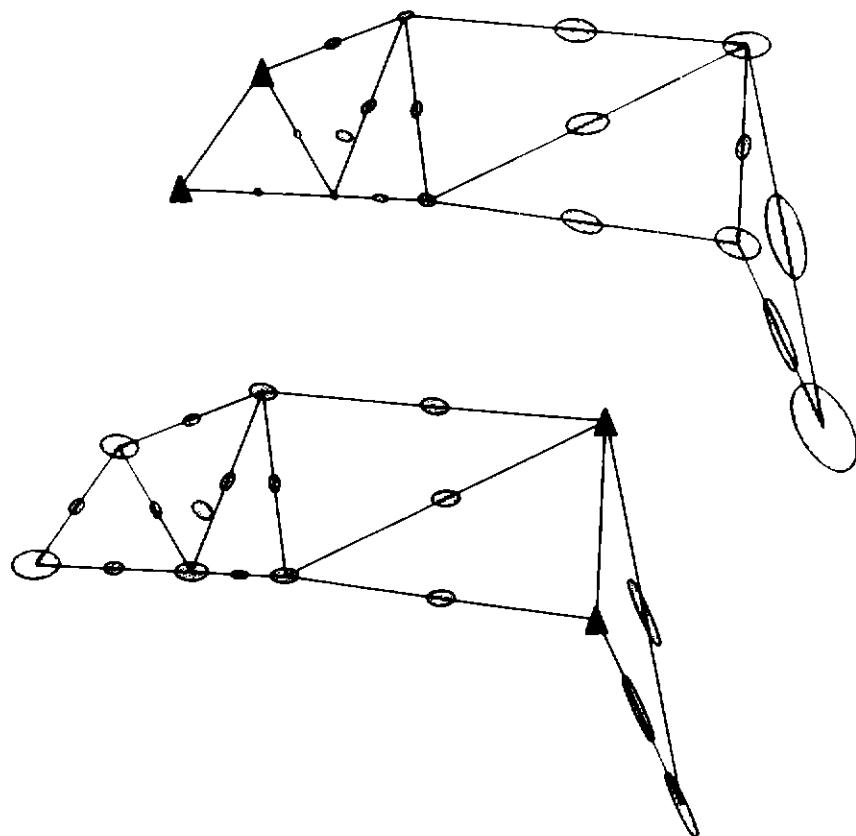


FIG. 18.10. Change of confidence ellipses with change of initial points ▲.

to three-dimensional as well as one-dimensional networks, discussed in the next chapter. They are based on the fact that the normal equations for the coordinate increments in regularly shaped networks have the same formal structure as a finite difference analogue of a partial differential equation of elliptic type with constant coefficients [BARTELME AND MEISSL, 1974]. Thus, techniques for solving partial differential equations can be used to obtain solutions, including the estimates of errors, for various *simulated two-dimensional networks*.

This approach is particularly appropriate for studying the behaviour of very large networks. For example, using this approach, the standard deviation  $\sigma_{\hat{l}}$  of an adjusted distance  $\hat{l}$  in an infinite equilateral trilateration network with distances measured to a standard deviation  $\sigma$  tends asymptotically to

$$\sigma_{\hat{l}} = \left( \frac{8}{3\sqrt{3}} \ln l + 0.699 \right)^{1/2} \sigma \quad (18.25)$$

[DUFOUR, 1970]. Both the effect of boundary control of a specific accuracy on the network and the effect of holes in the network can also be studied with this technique, if the network is viewed as a boundary value problem (cf. §3.2). Interested readers are referred to BORRE [1978].

There is a clear analogy between a trilateration network and a framework of hinged elastic bars. This analogy gives rise to the idea of using the equilibrium

conditions and elasticity equations of statics as another tool that gives further insight into the strength of horizontal networks. This approach is known as using the *abstract elasticity of the network*; for details see, e.g., HALMOS AND KÁDÁR [1977] and BORRE [1977]. Other ways to characterize the strength of horizontal networks include factor analysis [HARMAN, 1967] and contouring the covariance matrix [ALBERDA, 1974].

Recently, DARE AND VANÍČEK [1982] formulated a novel approach to strength analysis using the dictionary definition of strength as a measure of resistance to change. Their computational procedure is based on the philosophy that a network is as strong as its weakest link. They recognized three independent measures of strength: in scale, in twist (or in differential rotation), and in shear (for the definition of these strain-related terms, see §27.4).

Let us turn now to the assessment of *systematic distortion in horizontal networks*. By far the best way of studying these effects is to formulate them mathematically, if the role they play is known. We then get the numerical values that characterize the resultant distortions of the network. As an example, FIG. 11 shows distortions in a triangulation chain in Labrador (Canada) due to neglect of geoidal heights (a), and neglect of the deflections of the vertical in the reduction of observations to the horizontal datum (NAD 27) (b), as evaluated by THOMSON ET AL. [1974]. It is interesting to note that, in the first case, we have a simple scale distortion of  $-1.7 \times 10^{-6}$ , since only one distance in the chain was observed. The latter distortion is more irregular due to the irregularity of both the deflection field and the shape of the chain. It may be, nevertheless, approximated by a combination of a scale

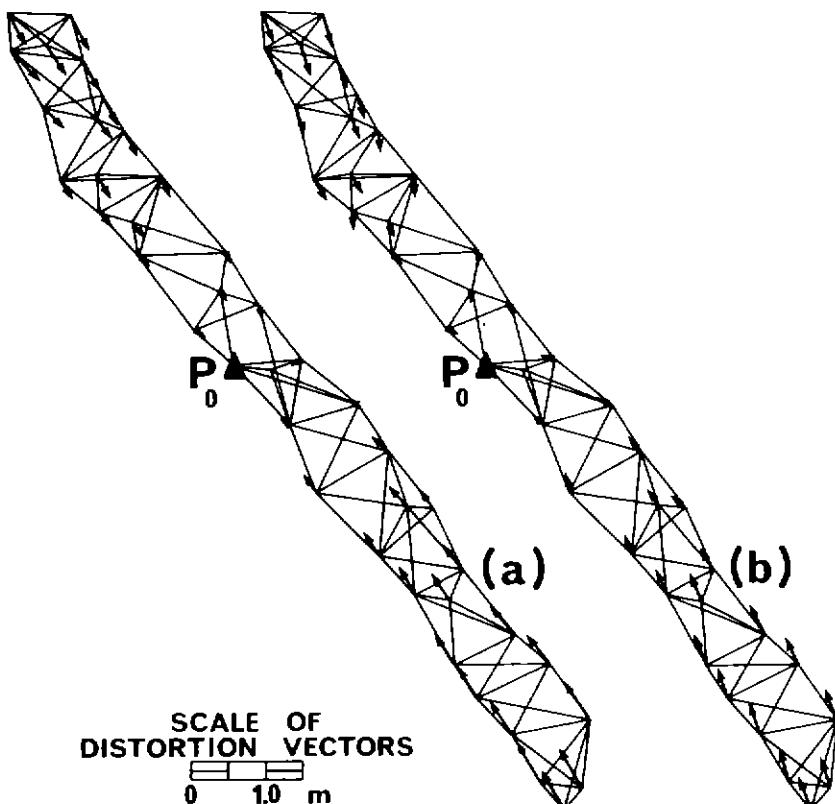


FIG. 18.11. Systematic distortions of a network due to (a) neglect of geoidal heights; (b) neglect of the deflections of the vertical.

distortion of  $-1.2 \times 10^{-6}$  and a misrotation of  $-0.065''$  around  $P_0$ . Interested readers are referred to MEISSL [1974] for an alternative approach to the analysis of systematic distortions that uses simulated networks of regular shapes.

If local distortions in the network are to be studied, it is advantageous to use the methods of differential geometry and portray the distortions as *network strain*. Since strain is going to be dealt with more naturally in §27.4 in the context of horizontal movements, it would be superfluous to go into detail in this section. Suffice it to say that the strain techniques as explained in §27.4 are applicable here as well. These techniques are particularly suitable in studying distortions induced by a single observation or a group of observations incompatible with the network, as defined by the rest of the observations [VANÍČEK ET AL., 1981].

We are now in a position where we can finally discuss the ways horizontal networks are expanded. By *expansion of a horizontal network* is meant the addition of new blocks of points to an existing, adjusted network. How can this problem be handled? Clearly, one can use the strategy employed in Helmert blocking as described in the previous section: the points of the existing network at which the new block is attached are regarded as junction points and dealt with accordingly. Naturally, the already adjusted coordinates of existing points change under the influence of newly admitted observations from the expansion network (see FIG. 12).

In geodetic practice, however, the change in the coordinate values of existing points, any time the network is expanded, is not looked upon favourably. Hence, other approaches are normally used of which only the most appropriate one will be discussed here. It consists of adding the reduced original normal equations (weight matrix) of only the junction points to the normal equations of the expansion network and finding the solution for the expansion network points including the junction points. It represents merely the generalized adjustment whose mathematical solution was shown in §14.4, and which was spelled out in terms of the main system of equations in Helmert blocking in §18.2.

This approach allows the random errors of the original network to propagate into the new network. On the other hand, the original coordinates of the junction points (in the original network) are not changed, if the new positions do not depart too much from the existing ones. The significance of such departures should be statistically tested using techniques shown in Chapter 13.

It is often the case, however, that the original positions are too distorted and

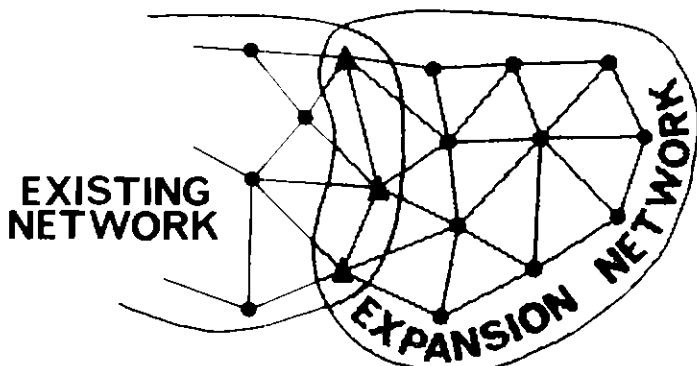


FIG. 18.12. Expansion of a horizontal network.

depart significantly from the newly determined positions. This is a consequence of either some systematic effects on observations, old or new, or the non-rigorous procedure that had been used in the modelling and adjustment of the original network. Under these circumstances, it makes sense to seek an *analytical model of the distortions* in terms of, for instance, a linear form. Least-squares regression is the natural technique to use in this context. The distortion can then be predicted for any existing point in the area and subtracted from the point coordinates to give the correct position.

Naturally, an expansion network can cover a new area or it can fill a gap in the existing network; the mathematical formulation remains the same. When the expansion is meant to fill a gap, we speak of the *densification of a horizontal network*. Historically, densification was carried out with observations of lower accuracy than those of the original, densified network. Nowadays, observations of the same, or even higher, accuracy may be used. In the United States, the readjustment of the horizontal networks is being done with data from networks of all orders [DRACUP, 1978].

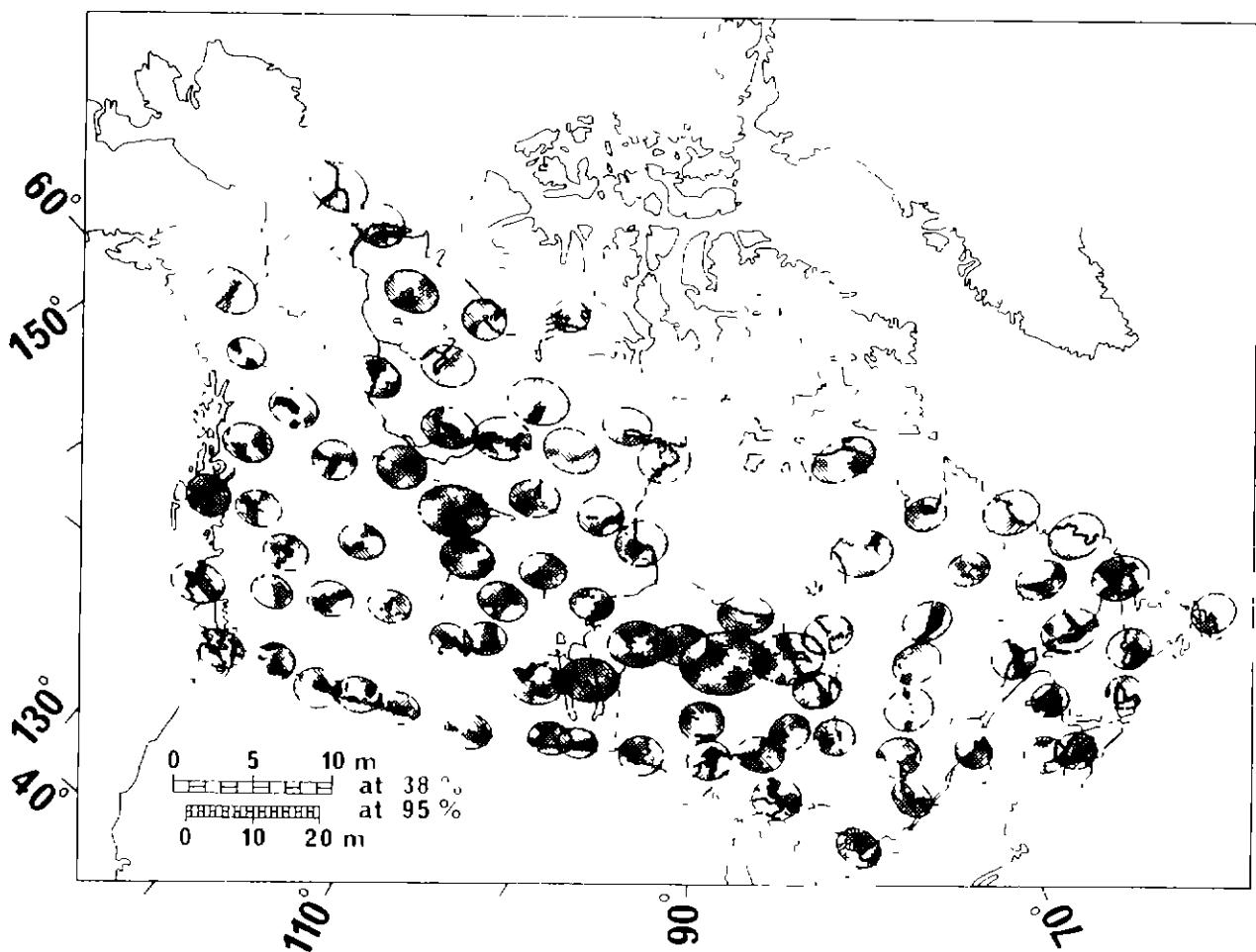


FIG. 18.13. Results of the merger of the TRANSIT satellite and terrestrial networks in Canada. (Courtesy of the Geodetic Survey of Canada, DEPARTMENT OF ENERGY, MINES AND RESOURCES [1979], Ottawa, Canada.)

The last topic to be discussed in this section is the *merger of horizontal networks* of different kinds that cover the same area. From the conceptual point of view, there is, of course, not much of a difference between merging two networks and expanding an old network: only now the junction points (common to both networks) are distributed throughout the old and new networks. The mathematical techniques used for merging horizontal networks are much the same as those discussed in §17.4 in the context of three-dimensional networks. The horizontal positions of points in both networks are expressed as three-dimensional coordinates that, for each network, happen to lie all on the appropriate reference ellipsoid (horizontal datum). After that, the problem is treated in three dimensions, with the constraint that the resulting positions must again be on an appropriate ellipsoid.

It should be pointed out that if one of the networks is a satellite network, where the horizontal positions have been determined by projecting the inherently three-dimensional coordinates onto a geocentric reference ellipsoid of a selected size and shape, then this reference ellipsoid can be enforced as a horizontal datum for the terrestrial network through the merger. The resulting position accuracies are of interest: with the position of the new horizontal datum being implied by the satellite-derived coordinates, the tendency for the absolute error ellipses to grow with distance from the initial point of the terrestrial network is largely checked. The central role of the initial point disappears. For illustration, see FIG. 13.

#### 18.4. Marine positioning

Up to this point, we have discussed positioning on land which normally implies the positioning of fixed, marked points. *Marine positioning*, i.e., positioning in the marine environment, has a different character. Disregarding points on the sea floor, there are very few *fixed objects* on the sea surface; these are confined to islands, rock formations, and the like. Usually these fixed points become increasingly scarce with distance from shore. The objects to be positioned are thus mostly moving all the time. These moving objects belong to one of the following two species:

- (a) objects anchored to the bottom, comprising buoys, drilling rigs, anchored vessels, etc.;
- (b) floating (steaming or drifting) objects such as vessels, icebergs, ice sheets, etc.

*Anchored objects* are normally bobbing, swinging, and moving irregularly within a fixed radius around some mean position, due to changing currents, waves, and other motions of the sea (cf. §8.4). Depending on the purpose, the mean position of the object may be of interest, or any of the instantaneous positions (see FIG. 14) are good enough, or, under special circumstances, a time series of positions may be required. *Floating objects*, on the other hand, follow some *course*, and it is usually of interest to determine the position of the object as a function of time to know where the object is at a specified instant of time. This is particularly true when one is concerned with the position of steaming ships where positioning is a substantial part of *navigation*. To realize how substantial a role positioning plays in navigation, it suffices to consult FIG. 15 illustrating the definition of navigation: navigation may

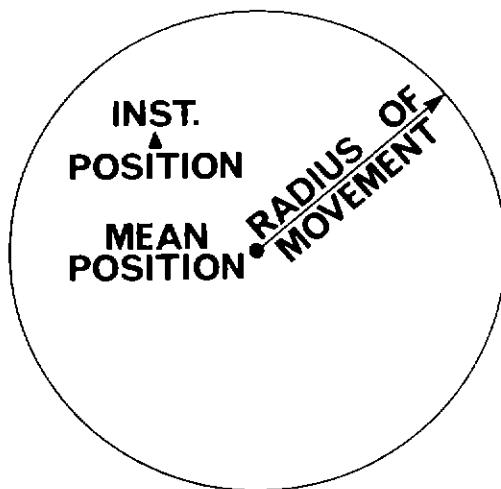


FIG. 18.14. Position of an anchored object.

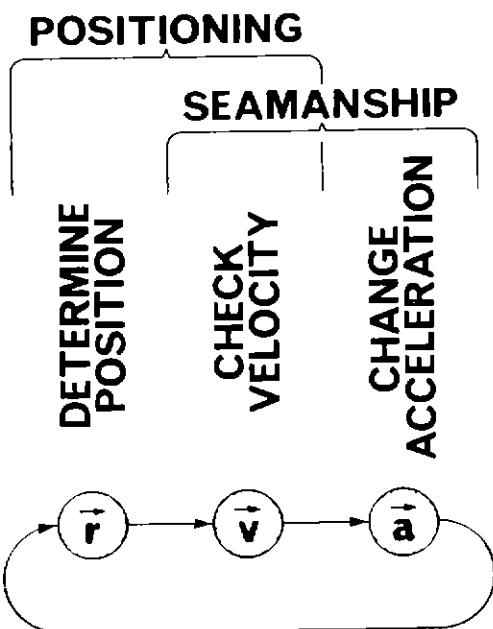


FIG. 18.15. Navigation.

be described as a feedback loop whose aim it is to direct the movement of a craft, expeditiously and safely, from one point to another [BOWDITCH, 1977].

There are some similarities and also some differences in the positioning of the three families of objects as defined above. Ordinarily, none of the terrestrial techniques based on direct intervisibility between points can be used. Further, only point positioning and relative positioning techniques can be employed; because of the dynamic nature of marine positioning, the network approach is out of the question. The main difference between positioning the fixed and anchored objects on the one hand and the floating objects on the other is that, in the former case, successive determinations of positions lead to redundancy and thus to the possibility

of estimating the accuracy of the determination. In the latter case, successive determinations, of necessity done in real time, give the course of the floating object as a function of time; unless redundant simultaneous observations are made, no redundancy arises, and there is no possibility of repeating the position determination. Also, because of the difference in the purpose of navigation, positioning done for navigation alone is done to a much lower accuracy standard. This may result in different methods being used in navigation.

In the context of marine positioning, one can speak of three-dimensional positioning when the depth is also to be determined. This is the case of hydrography, whose mandate is to chart the sea floor (cf. §7.1). The third dimension, however, can be treated separately. We will return to depth measurements in §19.4, where it can be dealt with more naturally. In this section, only horizontal positioning, i.e., positioning on the sea surface, will be treated. Positioning of points on the sea floor requires the transfer of horizontal positions from the sea surface down, and is considered to be beyond the scope of this book.

Focussing now on positioning on the sea surface, what are the differences between the observing and mathematical techniques used here and those used for horizontal positioning on land? Clearly, the horizontal position of fixed objects can be determined in exactly the same manner as that of the points on dry land. If they are not visible from any other fixed points located on shore or islands, extraterrestrial methods have to be used.

With anchored and floating objects, the situation is similar. If the object to be positioned is visible from at least two control points on the shore or islands, its position can be intersected as shown in FIG. 16. Either the horizontal angles  $\omega_{123}$  and  $\omega_{213}$  can be observed on shore or, what is much more often the case, the two ranges  $\Delta r_{13}$  and  $\Delta r_{23}$  may be determined. We shall show here how the *intersection mathematical model* for observed ranges is formulated on the reference ellipsoid. The formulation for observed angles can be found, for instance, in BOMFORD [1971].

First, some approximate coordinates  $\phi_3^{(0)}, \lambda_3^{(0)}$  of  $P_3$  are determined using a spherical or even plane approximation to the actual case. Then a linearized observation equation for spatial distances (4) is formulated for each observed range. If there

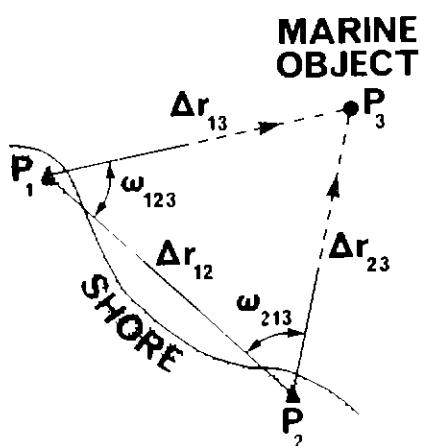


FIG. 18.16. Intersection of an object from two known points.

is no redundancy, one cannot seek the residuals  $r_{\Delta r}$  and also  $\delta\phi_1, \delta\lambda_1, \delta\phi_2, \delta\lambda_2$  must be put equal to zero. The resulting system of two linear equations for two unknowns  $\delta\phi = \delta\phi_3, \delta\lambda = \delta\lambda_3$  then reads

$$\begin{aligned} c_4(P_1, P_3) \delta\phi + c_5(P_1, P_3) \delta\lambda &= \Delta r_{13} - \Delta r_{13}^{(0)} = w_{13}, \\ c_4(P_2, P_3) \delta\phi + c_5(P_2, P_3) \delta\lambda &= \Delta r_{23} - \Delta r_{23}^{(0)} = w_{23}. \end{aligned} \quad (18.26)$$

where  $\Delta r^{(0)}$  are the spatial distances between  $P_1$  and  $P_3^{(0)}$ , and  $P_2$  and  $P_3^{(0)}$ . In the expressions for coefficients  $c_4$  and  $c_5$ , the heights  $h$  and vertical angles  $v$  can be taken as equal to zero and we get, after a little development,

$$\begin{aligned} \delta\phi &= \operatorname{cosec} \omega_{321}^{(0)} \left( \frac{\sin \alpha_{32}^{(0)}}{M_3^{(0)}} w_{13} - \frac{\cos \alpha_{32}^{(0)}}{N_3^{(0)} \cos \phi_3^{(0)}} w_{23} \right), \\ \delta\lambda &= \operatorname{cosec} \omega_{321}^{(0)} \left( \frac{-\sin \alpha_{31}^{(0)}}{M_3^{(0)}} w_{13} + \frac{\cos \alpha_{31}^{(0)}}{N_3^{(0)} \cos \phi_3^{(0)}} w_{23} \right). \end{aligned} \quad (18.27)$$

Here, all the quantities with superscript (0) are determined from the (approximate) coordinates of  $P_3^{(0)}$ . Naturally, if  $P_3^{(0)}$  is very far away from its correct position  $P_3$ , then these quantities may have to be changed in the next iteration. In marine positioning, the aforementioned approach is called the *range-range positioning*. When more than two ranges are observed, the mathematical model becomes over-determined and the least-squares solution is sought.

If neither onshore angles nor ranges can be measured and have to be replaced by angles measured at sea, then the situation becomes more complicated. The first problem is the lack of a firm support for the angle measuring instrument: the observing station at sea moves constantly. It either oscillates by a few metres (as on a drilling platform) or progresses on the course up to several metres per second (on a steaming ship), and it may yaw, roll, and pitch with the vessel. The observations are thus usually carried out with a sextant, a hand-held instrument [ANDERSON, 1966]. If higher accuracy is essential, then stabilized platforms (see, e.g., VON ARX [1967]) can be used. It is also necessary to see at least three onshore control points and then the mathematical model becomes more sophisticated. We speak of a *resection mathematical model* [CLARK, 1969], which we are not going to show here.

Another alternative often used in marine positioning is to measure range differences. It can be shown that on a plane the locus of all the points that have the same range difference from two known points  $P_1, P_2$  (see FIG. 17) is a hyperbola. Clearly, if the range difference is also measured to another pair of known points (e.g.,  $P_1$  and  $P_3$  in FIG. 17), then another hyperbola can be plotted, and the positioned object is located at the intersection of the two hyperbolae. This can be expressed mathematically, and one obtains the *mathematical model for hyperbolic positioning* [INGHAM, 1974].

A significant proportion of marine positioning is done at distances much longer than the line-of-sight distances encountered on land. This rules out the use of very

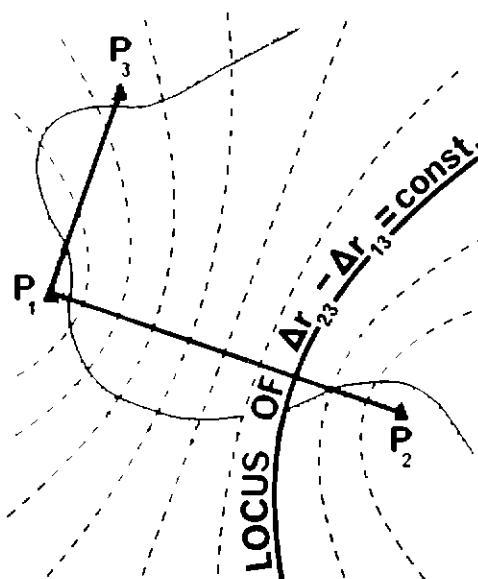


FIG. 18.17. Measurement of range differences.

high and higher (including visible) frequencies for other than close to shore positioning, and high to low frequencies with their surface propagation mode have to be used instead (cf. §9.2). The existing medium range systems, such as Hi-Fix, Argo, and Raydist, use medium frequencies, the long range systems, such as Decca and Loran-C, use low frequencies; and the global range Omega system uses very low frequencies [THOMSON AND WELLS, 1977]. As a result of the longer distances involved, the accuracy of the range and range difference determination is significantly lower than that on land. The relative accuracy of these systems is typically of the order of  $10^{-4}$  compared with  $10^{-5}$  to  $10^{-6}$  on land. It is interesting to record that in spite of this lower accuracy, long range positioning (Shoran, Aerodist) was used in Canada a few decades back to establish the first geodetic networks in inaccessible regions [DEPARTMENT OF MINES AND TECHNICAL SURVEYS, 1955].

The accuracy of a measuring system generally degrades with distance from the shore. It can be conveniently argued, at least in some instances, that the requirements for accuracy also become less stringent in deeper water and thus further from shore. Consequently, it is reasonable to divide position accuracy requirements into zones; one such division is shown in TABLE 2 [THOMSON AND WELLS, 1977].

TABLE 18.2  
Positioning zones

Zone	Range	Representative depth
Inshore (harbours, rivers, and estuaries)	0–30 km	40 m
Coastal (bays, inlets, and straits)	30–150 km	100 m
Offshore (to the edge of the continental shelf)	150–1000 km	200 m
Deep sea	> 1000 km	1000 m

The accuracy of a marine position depends on the instrumental accuracy, knowledge of the propagation velocity, influence of the environmental factors, geometry of the position fix configuration, elimination of systematic errors, and the appropriateness of the mathematical models used for the position computations. Just as the marine environment dictates that both instrumentation and observing methodology be different than on land, it also affects the treatment of observations. For direction and angle measurements made from the shore to a surface vessel, none of the corrections discussed in §16.2 is applied, since the corrections themselves are very small relative to the standard deviations of the measurements. Angular measurements made from a vessel to onshore points are corrected only for the skew-normal when the vertical angle of the target exceeds one degree of arc. Electromagnetic range and range difference measurements are subjected to the same treatment as on land, although, in many instances, coarser approximations are made since the corrections are again substantially smaller than the standard deviations of the measurement.

The effect of tropospheric refractivity (see §9.2), called the *primary phase lag* here, is simply to slow down the signal. It can be accounted for by assuming constant refractivity over the propagation path. The effect on propagation of the conductivity of the earth's surface, the curvature of the earth, and other phenomena are much more complicated [BREMNER, 1949; JOHLER ET AL., 1956]. Their combined effect is treated as the problem of *secondary phase lag* correction. Considerable research has been done on measuring and predicting the phase of ground waves over homogeneous and mixed land-water propagation paths, e.g., BRUNAVS AND WELLS [1971].

The use of medium to low frequencies for measuring distances beyond the optical line-of-sight introduces a new factor. In considering the propagation of surface waves over the earth's surface, the conductivity and permittivity of the earth's surface layers are important. The ground acts both as a capacitor and resistor in parallel carrying electrical currents induced by surface waves. These currents are attenuated with depth; the depth of penetration decreases with frequency and conductivity. Because the earth's surface is an imperfect conductor, the ground edge of the wave-fronts on the surface will be retarded, causing the wave to tilt forward (FIG. 18). This tilt and decreases with frequency and conductivity. Over sea water it is never more than a few degrees; over land it may exceed 20 to 30 degrees. This

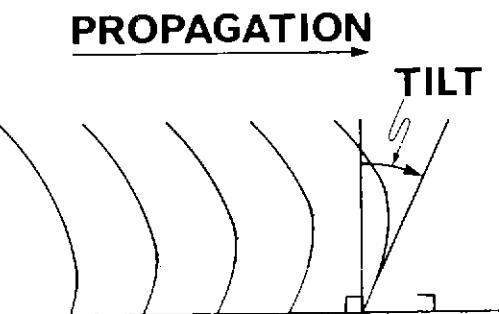


FIG. 18.18. Tilt of radio waves due to retardation by the earth's surface.

retardation introduces a *phase lag of the surface wave* which increases with distance from the transmitter and decreases with conductivity and frequency.

There exist many more techniques for positioning at sea based on other principles. The most successful short range systems use sea floor control. Since, as we saw in §9.2, the electromagnetic waves do not penetrate into water very well, sound waves are used instead, referred to as *acoustic positioning systems* [MACPHEE, 1976]. Another family of positioning devices are those based on sensing differences in the horizontal velocity or in the horizontal acceleration (or in both of them) of the vessel; these are used almost exclusively for navigation purposes. The sensors are either inertial (cf. §16.1), magnetic, or those sensing the instantaneous velocity of the vessel with respect to the surrounding water. These systems are known as *self-contained positioning systems*, and their description can be found in e.g., BOWDITCH [1977].

From the extraterrestrial techniques used in marine positioning, one should mention the *astronomical determination of marine position*, a family of the oldest, and still extensively used, techniques known to navigators. These astronomical techniques exploit the relative ease with which zenith distances can be measured with a sextant and time with a chronometer. Any of the mathematical models shown in §15.2 can be used here [BOWDITCH, 1977] even though the accuracy of observations is low. If more accurate results are desired, then a universal theodolite mounted on a stabilized platform should be used [VON ARX, 1967].

Of the satellite positioning systems, only the range difference (TRANSIT) system, originally conceived for marine positioning, is now being used. The measuring equipment needed for the other techniques (e.g., satellite cameras, lasers, dish-like antennas) is too bulky or fragile to stand the rigours of a marine environment. Even the TRANSIT system (see §15.3), however, cannot be employed to its fullest because of the motion of the vessel. This motion interferes with the satellite motion Doppler effect by adding a spurious component to it. It also prevents the use of more than one satellite pass for the position determination if the ship steams on a course. Taking all sources of error into account, EATON ET AL. [1976] estimate that a TRANSIT satellite offshore position accuracy can vary between 60 and 600 metres.

The situation is going to change rapidly once the full capability of the NAVSTAR Global Positioning System becomes available (see §15.3). GPS, having been designed primarily as a navigation tool, will naturally take over from the TRANSIT system to which it is vastly superior, particularly in the marine environment. The prospects for GPS have been analysed by WELLS ET AL. [1982].

If the expense is justifiable, a combination of measuring systems can be used that is capable of producing more than a sufficient number and kind of observations for a unique position determination. Then, adjustments and statistical analyses can be carried out. When there are at least two different, complementary positioning systems employed in the dynamic (navigation) mode, the technique of Kalman filtering (see §14.6) can be applied. This approach is being used with the more recently developed integrated navigation and positioning systems such as BIONAV [WELLS AND GRANT, 1977].

## CHAPTER 19

# HEIGHT NETWORKS

Although both the measuring procedure and the treatment of height networks are conceptually more simple than was the case with the horizontal or three-dimensional networks, to get the accuracy of the heights of network points inherent in geodetic levelling requires a good understanding of the physical processes involved. This is because heights determined from geodetic levelling are more intimately dependent on the earth's gravity field than are horizontal positions. Also, atmospheric refraction plays a more important role in levelling than it does in horizontal positioning. Both these effects accumulate in a more or less systematic manner. All these problems are treated in detail in this chapter, which expands on the introductory description of vertical networks in §7.1.

Another problem, of a purely physical nature, is the realization of the vertical datum; it is treated in the first section. The second section is devoted to mathematical models used in height networks and to the nature of systematic and random errors in levelling. The third section deals with the concepts involved in the assessment, design, and densification of height networks. The last section is only obliquely related to height networks: it deals with height determination along profiles in different environments using different concepts.

### 19.1. Vertical datum

As stated in §7.1 and §16.4, the orthometric and, with a qualification, the dynamic heights use the geoid as the vertical datum, while the datum for normal heights is the quasigeoid. It was also shown that, disregarding the less precise techniques (e.g., trigonometrical levelling), it is the levelled height differences that are measurable and convertible into orthometric, dynamic, or normal height differences through the application of appropriate corrections (cf. §16.4). The question then is: How can one realize the zero reference for these heights? The usual way of obtaining heights from the appropriate height differences is to begin at the sea shore where the geoid, or quasigeoid, are accessible. On the oceans, the quasigeoid and geoid coincide (cf. §7.4), thus, if we locate one of them at a point on the shoreline, we have automatically located the other. For the sake of simplicity, we shall talk about only the geoid here.

Until a few decades ago, it was tacitly believed that the mean sea level (see §7.2) should theoretically coincide with the geoid, or that the departure of the two surfaces was (and still is) negligible. Hence, the task of locating the vertical position of the vertical datum, i.e., the geoid, with respect to a *reference bench mark* on the shore was reduced to the task of determining the position of the mean sea level. For this purpose, the variation of the *local instantaneous sea level*  $H_{ISL}$  (with respect to the zero of the recording tide gauge) was, and still is, being recorded. The *local mean sea level*  $H_{MSL}$  is calculated, and the height of the reference bench mark above the mean sea level  $H_{MSL} + \Delta H_{BM-TG}$  is established, as depicted in FIG. 1. The heights of all the other points in the network are then obtained from the heights of these reference bench marks by accumulating the height differences along the interconnecting levelling lines (see, e.g., CANNON [1929]).

Although heights above the local mean sea level are still used all over the world, as we now know they are only approximately equal to the proper heights above the geoid (quasigeoid). The difference is caused by the sea surface topography (cf. FIG. 1) discussed in §7.2. Its magnitude may easily amount to several decimetres. By forcing the height of the local mean sea level to zero, i.e., by neglecting the sea surface topography, we warp the height network causing the heights of all the network points to be distorted.

There is another problem inherent in the determination of the local mean sea level: it is generally dangerous to evaluate the mean level by taking the straight average of the sample. This may result in a bias; an illustration of this phenomenon is shown in FIG. 2. Thus a more involved treatment is recommended. To appreciate such a treatment, we have to take a closer look at the causes of the actual variation of sea level. The short periodic variations due to tsunamis, waves, semidiurnal and diurnal tides, as discussed in §8.4, can be effectively filtered out using a low-pass

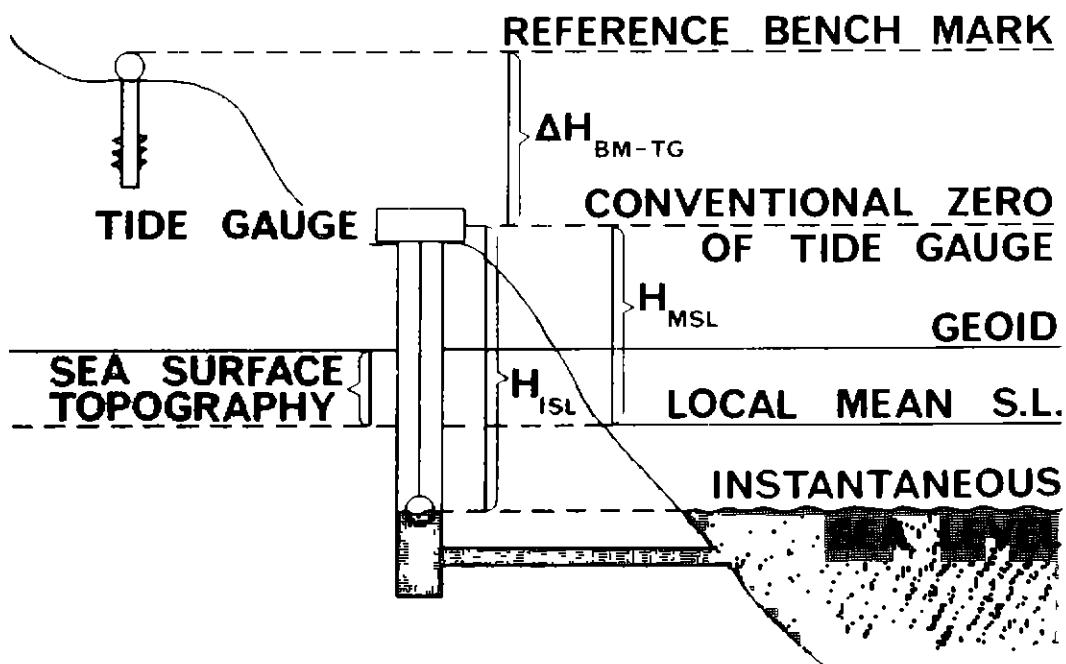


FIG. 19.1. Establishment of height of reference bench mark.

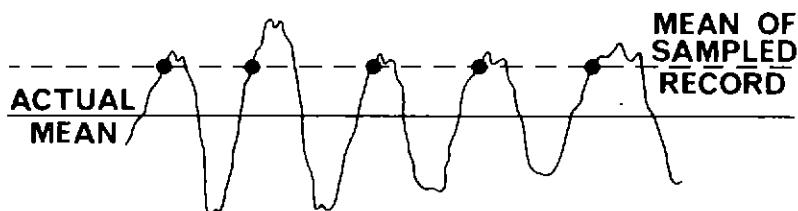


FIG. 19.2. Example of aliasing.

filter (cf. §14.2). More troublesome are secular and long periodic effects which affect the monthly or annual means normally used for the determination of the local mean sea level.

MONTGOMERY [1937–38] recognized the following long periodic causes of sea level oscillations: (a) atmospheric pressure variations, (b) dynamic effects of sea current changes, (c) wind variations, (d) thermohaline changes, (e) river discharge fluctuations, (f) changes in bathymetric configurations, (g) glacial melt, and (h) long periodic tides. In addition, there is a periodic variation induced by polar motion that will be explained in §25.4.

(a) The changes due to *atmospheric pressure* variations may, under special conditions, reach up to several decimetres [RODEN, 1966]. The larger the pressure the larger the depression; the coefficient of proportionality varies widely with location around the value of 1 centimetre per millibar.

(b) The time variations of *sea currents* are little known. As we shall see later in this section, it is difficult to get even the data for the determination of the first-order stationary term responsible for the sea surface topography. This means that, even more so than for time variations, it is next to impossible to model the higher order terms.

(c) The *wind stress* effect has been studied extensively, mainly because of the threat to life and property from its short periodic variety—the storm surges that may amount to several metres [MILLER, 1958]. The long periodic effects are, of course, much smaller. For example, for the port of Halifax, Canada, ANDERSON [1978] reported that the effect of the wind component normal to the shore (piling-up effect) has a maximum of a few decimetres in monthly averages.

(d) *Thermohaline* (thermal and solution) *structure* of sea water is probably among the most important causes of long periodic changes in sea level. Fortunately, it is generally stable (stationary) with superimposed seasonal variations of only thermal origin in the uppermost layers. RODEN [1966] computed the thermal effect to be between 1 and 3 centimetres per degree Celsius along the western coast of the United States. Little is known about temporal variations in salinity.

(e) *River discharge* fluctuations can contribute significantly to long term variations, with the degree of significance heavily dependent on location. The contribution may reach the decimetre range [VANÍČEK, 1978]. MEADE AND EMERY [1971] estimated the discharge along the coast of the eastern United States to account for 7%–21% of the sea level variability.

(f) In comparison, the effect of *sea bed* is extremely difficult to quantify without numerical modelling.

(g) *Glacial melt* and the yield of the earth to the melt load are presumed to be the main constituents of the secular water rise mentioned in §8.4. The actual value of secular rise is still a matter of considerable uncertainty; its estimates vary between 6 and 10 centimetres per century.

(h) The magnitude of the *long periodic tidal constituents* is so small as to be of little practical consequence. The annual tide has a theoretical amplitude of about 0.5 cm [ROSSITER, 1966]; the actual amplitude cannot be determined very well because of aliasing with the temperature effect. The equilibrium value of semi-annual tidal amplitude is about 3 cm [ROSSITER, 1966] and, again, its actual value is unknown due to interference from meteorological effects. Motions of the lunar perigee (period of 8.85 years) and node (period of 18.6 years) produce variations of the order of 1 cm [VANÍČEK, 1978]. It is interesting to note that it was originally thought necessary to have at least 18.6 years of sea level data available to eliminate the influence of the nodal cycle. It now appears that this precaution is no longer justified because of the small amplitude of the 18.6-year constituent.

Two more phenomena should be mentioned: the variations with the Chandler period (see §5.4) are detectable [CURRIE, 1975] but insignificantly small; and there are vertical crustal movements of a local and regional nature that affect the tide gauge readings. These will be discussed in detail in Chapter 26; let it suffice to say here that the effect of crustal movements on sea level records should, in the present application, be eliminated if at all possible.

We are finally in a position to discuss a possible linear filter for obtaining the value of the local mean sea level. Denoting the  $n$  available time series of meteorological and other auxiliary data (water temperature, barometric pressure, etc., recorded for the location of the tide gauge) by  $P_i(\tau)$ ,  $i = 1, \dots, n$ , we can write the following linear filter:

$$H_{\text{MSL}}(\tau_0) = l(\tau) - c_E(\tau - \tau_0) - \sum_{i=1}^n c_i (P_i(\tau) - \text{mean } P_i) - \sum_{j=1}^5 (a_j \cos \omega_j \tau + b_j \sin \omega_j \tau). \quad (19.1)$$

Here,  $l(\tau)$  is the sea level record discretized by taking individual values of  $\tau$  (a time series, consisting of, e.g., monthly averages),  $\tau_0$  is the epoch for which the value of the local mean sea level is sought,  $c_E$  is the rate of secular water rise combined with local linear crustal movements,  $\omega_j$  ( $j = 1, \dots, 5$ ) are the frequencies of long periodic tidal constituents and the Chandler frequency discussed earlier. Coefficients  $c_E$ ,  $c_i$  ( $i = 1, \dots, n$ ),  $a_j$ ,  $b_j$  ( $j = 1, \dots, 5$ ), and  $H_{\text{MSL}}(\tau_0)$  can be determined from least-squares regression, as discussed in §14.2.

From the point of view of the task at hand, only the constant  $H_{\text{MSL}}(\tau_0)$  is the signal; the rest of the terms are a noise and do not have to be evaluated at all. It cannot be overemphasized that the value of  $H_{\text{MSL}}$  depends on the selected epoch  $\tau_0$ . To illustrate this point, FIG. 3 shows the sea level record from the port of Halifax, Canada, together with the residuals created by filtering out the noise due to linear trend, sea water temperature, barometric pressure, river discharge, wind piling,

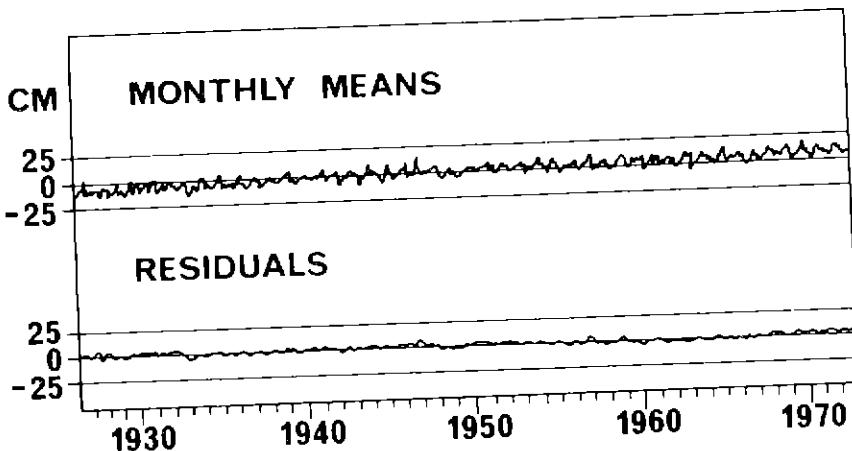


FIG. 19.3. Sea level record from the Port of Halifax, Canada. (Courtesy of ENVIRONMENT CANADA [1979], Ottawa, Canada.)

and the five periodic constituents [ANDERSON, 1978]. The mean square error in Anderson's determination of  $H_{MSL}$  (for  $\tau_0 = 1926$ ) was 0.3 centimetres.

There is one additional problem related to the eustatic water rise stemming from the definition of the geoid used in this chapter. Evidently, if the mean sea level is understood to be varying with time, then so does the geoid and the value of the actual potential defining it. This would be an awkward concept to live with in the context of heights. Therefore, the mean sea level and the vertical datum should be taken as constant for the period of time covering the life span of the vertical network [CASTLE AND VANÍČEK, 1980].

Of the three problems discussed so far—sea surface topography, time variation of MSL, and time variation of the geoid—the first is the most serious and, unfortunately, cannot be solved completely satisfactorily with our present knowledge of the actual behaviour of the oceans. There exist four conceptually different approaches for determining the sea surface topography: steric levelling, the study of global circulation, satellite altimetry, and the local response technique. *Steric levelling* is a technique whereby the variations of sea water density with depth are directly measured from a ship at a number of points. Then, based on a disputable assumption that horizontal ocean currents cease at depths of more than a few kilometres, the sea surface topography can be evaluated directly by integration over the sampled density profile at each point [BJERKNES AND SANDSTRÖM, 1910]. A *global circulation* pattern can be obtained by measuring the three-dimensional, time-varying distribution of water velocity. Based on these data, the differential equations of motion of the sea water can be formulated and solved. The sea surface topography is obtained as a by-product of such a solution. This technique was used, e.g., by HELA AND LISITZIN [1967] in their investigation already depicted in FIG. 7.8. Satellite altimetry, the only approach that uses modern spatial techniques, will be described in §19.4.

The first three approaches suffer from a common drawback: they give only the departures of the instantaneous sea level from the geoid, i.e., the *instantaneous sea surface topography*. The sea surface topography representative of a longer time

period can be obtained only after the observations and calculations have been repeated many times over that period of time. Moreover, the accuracy of the first two (oceanographic) techniques decreases toward the coast; steric levelling cannot be used in shallow waters at all. Yet the close offshore areas are the ones of particular interest to geodesy.

The fourth approach, the *local response technique*, is based on the idea of seeking a zero-frequency (permanent) response of local sea level to various local effects, such as temperature, atmospheric pressure, river discharge, wind stress, etc. Using the computed local anomalies of these effects (with respect to their global means), one gets the local value of sea surface topography. MERRY AND VANÍČEK [1983] report encouraging results for eastern Canada obtained from this technique.

It is easy to see why the sea surface topography has been deemed unknown and has not been corrected for in the establishment of vertical datums. An alternative to the above described standard approach that relies on tide gauges appears to be to take one point, somewhere in the middle of the network, as the *origin of the height network*. Then, heights of the local mean sea levels at all the connected tide gauges may be derived. Finally the heights of all the points, including the origin and all the reference bench marks, can be corrected by a constant amount determined in such a way as to make the local mean sea level heights look realistic everywhere. This alternative should be selected only when the uncertainty in the sea surface topography is too large for the heights of the reference bench marks to be useful as weighted constraints in the adjustment.

The last point worth discussing is the appropriateness of the geoid versus the quasigeoid as a vertical datum or, in other words, the appropriateness of orthometric versus normal heights. In any practical mapping or surveying application, all that is actually needed are height differences, of one kind or another, on the earth's surface; it does not matter at all where the datum is located underground. In addition, it should be remembered that even these height differences (orthometric or normal) have no physical meaning anyway, hence a case cannot be made for any preference. The only application where the choice of the datum does matter is when the heights, together with horizontal positions, are to be converted into three-dimensional coordinates, or vice versa. In this situation, either geoidal heights  $N$  or height anomalies  $\zeta$  should be known, as was seen in §15.4.

## 19.2. Mathematical models for levelling

It was shown in (16.87) that the integral of geopotential numbers around a closed loop is zero. Clearly, the same holds true for dynamic heights, where we have

$$\oint_C dH^D = \frac{1}{g_R} \oint_C dC = 0. \quad (19.2)$$

For any orthometric height, one again obtains

$$\oint_C dH^O = 0. \quad (19.3)$$

To prove this, let us first write

$$H_i^O = H_i^D \frac{g_R}{\bar{g}_i'} = H_i^D + H_i^D \frac{g_R - \bar{g}_i'}{\bar{g}_i'}. \quad (19.4)$$

Writing a similar equation for  $H_j^O$  and subtracting (4) from it, we get (cf. (16.92))

$$\Delta H_{ij}^O = \Delta H_{ij}^D + H_j^D \frac{g_R - \bar{g}_j'}{\bar{g}_j'} - H_i^D \frac{g_R - \bar{g}_i'}{\bar{g}_i'} = \Delta h_{ij} + OC_{ij}, \quad (19.5)$$

where  $OC_{ij}$  is called the *orthometric correction*. Further, let us consider, without any loss of generality, the loop  $\mathcal{C}$  to be composed of two lines as shown in FIG. 4. Then one gets

$$\oint_{\mathcal{C}} dH^O = \Delta H_{ij}^O + \Delta H_{ji}^O; \quad (19.6)$$

substitution for orthometric height differences from (5) results in

$$\oint_{\mathcal{C}} dH^O = \Delta H_{ij}^D + \Delta H_{ji}^D, \quad (19.7)$$

since the other four terms cancel out. This equals to zero because the integral over  $\mathcal{C}$  of dynamic height differentials equals to zero (cf. (2)). It should be obvious that one can also speak about normal height differences and the *normal correction* to levelled height differences. It is left to the reader to derive the expression for the latter simply by substituting  $\bar{y}$  for  $\bar{g}'$  in (5). Similarly, the reader should be able to prove that

$$\oint_{\mathcal{C}} dH^N = 0, \quad (19.8)$$

by following the same reasoning as that for orthometric height.

Any height system can then be used in the *adjustment of levelling networks*, if the height differences are evaluated using gravity observed on the earth's surface. This is because the summation of height differences in any of these systems around a closed loop theoretically goes to zero, and this condition can then be used as the basis for the adjustment. If actual gravity on the earth's surface is not known, and normal gravity  $\gamma$  has to be used instead (cf. §16.4), heights defined in this way become path dependent, and their differences generally do not sum up to zero around a closed loop.

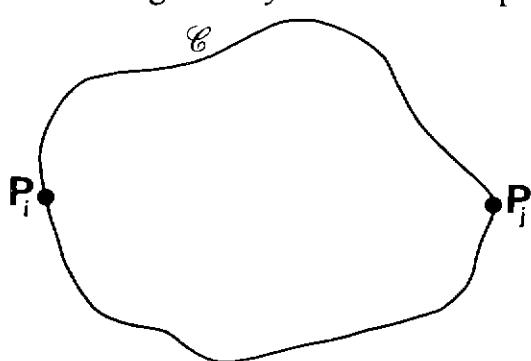


FIG. 19.4. Closed levelling loop.

The magnitude of the deviation between a height difference based on actual and normal gravity (approximate height difference) depends on how much the actual gravity  $g$  departs from  $\gamma$ , and on either the height difference or, for orthometric and normal heights, on the average height of the two end points of the segment. The formulae for these differences, or *gravity corrections* as they are called, for the approximate height systems used in North America have been derived by NASSAR AND VANIČEK [1975]. An example of the gravity correction values within a *levelling line* is shown in FIG. 5. Note that it does not make sense to talk about gravity corrections to approximate heights (only to approximate height difference along a levelling segment) because of their path dependency.

The question now arises as to whether the gravity corrections should be systematically applied to the approximate height differences. The answer must be 'yes', if the accuracy of the observed elevation differences is not to suffer from the conversion to height differences. Even though the gravity corrections behave in a random fashion over very long runs, this is certainly not the case for medium distances of tens and hundreds of kilometres. Over these medium distances, several investigators have confirmed that the discrepancies successfully compete for predominance with cumulative random errors in levelling (e.g., VIGNAL AND KUKKAMÄKI [1954], KRAKIWSKY AND MUELLER [1966]).

In the past few decades, the gravity data coverage on some continents has become sufficiently dense so that gravity values along the levelling route can be obtained for most areas through interpolation from nearby gravity observations. It has been shown [VANIČEK ET AL., 1980] that, to evaluate the gravity corrections for dynamic and Vignal heights to such an accuracy that their standard deviations are smaller than the corrections themselves (in absolute value), it is necessary to ensure that the

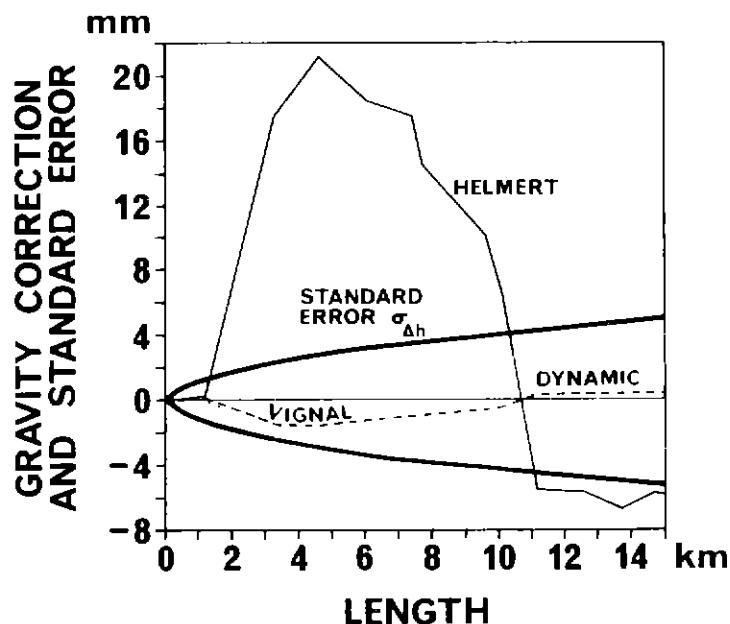


FIG. 19.5. Behaviour of gravity corrections along a levelling line in Alberta, Canada [NASSAR, 1977].

standard deviations of the free-air anomalies used satisfy the following inequality:

$$\sigma_{\Delta g} < |\Delta g|. \quad (19.9)$$

In the case of Helmert heights, we arrive at a more complicated formula: namely,

$$\sigma_{\Delta g} < \frac{|\Delta g_B - \Delta g_A - 0.2238 \Delta h_{AB}|}{\sqrt{2(1 - \text{cov}(\Delta g))}}. \quad (19.10)$$

Since, in the latter case, the key quantity is the horizontal gradient of gravity  $\nabla_s g$  along the segment  $S$ , (10) can also be rewritten as

$$\sigma_{\nabla_s g} < |\nabla_s g - 0.2238 \beta|, \quad (19.11)$$

where 0.2238 is in  $\text{mGal m}^{-1}$ , and  $\beta$  is the slope of terrain.

Equally as significant as the effect of the neglect of the actual gravity field is the effect of *residual refraction* (see §16.4). Its origin is in the fact that the air temperature stratification is irregular, as seen in FIG. 6. According to KUKKAMÄKI [1938], the *levelling refraction correction*  $\delta H_R$  for this effect is given as

$$\delta H_R = A \Delta t \Delta S^2 \delta l.$$

(19.12)

Here,  $\delta l$  is the observed elevation difference in metres,  $\Delta S$  is the sight length in metres, and  $\Delta t$  is the temperature difference in degree Celsius at two chosen elevations  $z_1, z_2$  above the ground. Further,

$$A = \frac{4.76 \times 10^{-4}}{z_1^c - z_2^c} \left[ \frac{-1}{1 + c} (z_1^{c+1} - z_2^{c+1}) + z_0^c (z_1 - z_2) \right] \quad (19.13)$$

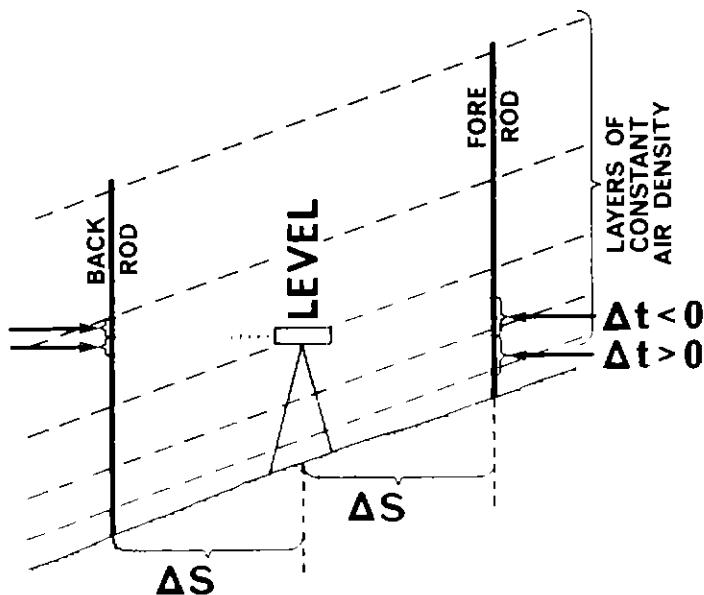


FIG. 19.6. Character of residual refraction.

in millimetres per metre cubed and per degree Celsius for  $z_0$ ,  $z_1$ ,  $z_2$  in metres. Here  $z_0$  is the elevation of the instrument above the ground, and  $c$  is the exponent in the usually assumed relation

$$t = a + bz^c \quad (19.14)$$

between temperature and elevation above ground. Typically,  $c$  is taken as equal to  $-1/3$  during the daytime so that, for  $z_1 = 0.5$  m,  $z_0 = 1.5$  m, and  $z_2 = 2.5$  m,  $A$  comes out to be  $-6.46 \times 10^{-5}$  mm/(m<sup>3</sup> °C). For  $\Delta S = 50$  m,  $\delta l = 2$  m, and  $\Delta t = -0.25$  °C, all fairly representative values, one obtains  $\delta H_R = 0.08$  millimetres. The main problem with applying this correction is in obtaining the proper value for the vertical increment of temperature  $\Delta t$ . It can be either measured in the field or estimated from known meteorological data—see, e.g., HOLDAHL [1980]. Alternative formulae to Kukkamäki's (12) exist—see, e.g., BRUNNER [1980].

Corrections for the thermal expansion and irregular graduation of rods are also being applied. Since these belong to the measuring process, they are not dealt with here; the interested reader is referred to, e.g., RAPPLEYE [1948] and BOMFORD [1971]. Tidal and tidal loading corrections will be dealt with in Chapter 25. Corrections for other crustal deformations should also be applied if the sources are sufficiently well known; regrettably, this is seldom the case.

Turning now to the mathematical model for the adjustment itself, it can be formulated either as a parametric or a condition case. We shall show here only the parametric approach, leaving the condition model for the reader to develop. The *observation equation for a levelling line connecting adjacent junction points*—e.g.,  $P_i$ ,  $P_j$  in FIG. 7—is written as

$$r_{ij}^{\Delta H} = H_j - H_i - \Delta H_{ij}^{(0)}, \quad (19.15)$$

where  $\Delta H_{ij}^{(0)}$  is the observed value of height difference in any of the height systems shown in §16.4, corrected for all the effects discussed above.

To adjust these observation equations, we have to know the covariance matrix  $C_{\Delta H}$  of the observed  $\Delta H$ , which we shall discuss first. To study this covariance matrix, let us denote the height difference between two adjacent running bench marks by  $\delta H$ ; this situation is shown in FIG. 8. The standard practice is that  $\Delta H$  is weighted

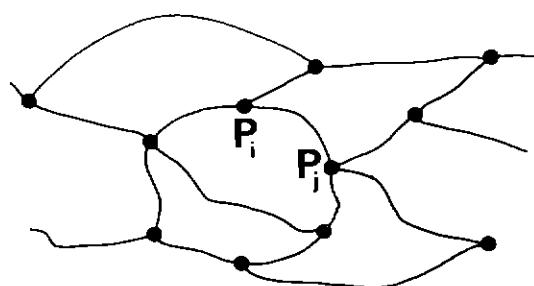


FIG. 19.7. Typical levelling network.

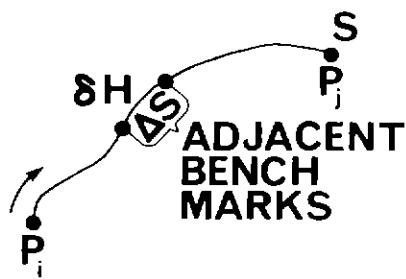


FIG. 19.8. Levelling line.

inversely proportional to the length  $S$  of the levelling line associated with the height difference. This weighting scheme is justified only when the height differences  $\delta H$  between the end points of individual *segments* (i.e., the connection between two consecutive adjacent bench marks) are statistically independent. Under this condition, the variance of  $\Delta H$  becomes

$$\sigma_{\Delta H}^2 = \sum_{i=1}^n \sigma_{\delta H_i}^2. \quad (19.16)$$

If all the height differences  $\delta H_i$  in a line are measured to the same accuracy, it is expedient to standardize all the variances  $\sigma_{\delta H_i}^2$  by expressing them in terms of the variance of a levelled height difference along a unit distance ( $\Delta S = 1$ ). The unit variance is denoted as  $\sigma_1^2$  and is evaluated normally for a distance of 1 kilometre. Under the assumption of statistical independence,

$$\sigma_{\delta H}^2 = \sigma_1^2 \Delta S, \quad (19.17)$$

and thus (16) becomes

$$\sigma_{\Delta H}^2 = \sigma_1^2 \sum_{i=1}^n \Delta S_i = \sigma_1^2 S, \quad (19.18)$$

or

$\sigma_{\Delta H} = \sigma_1 \sqrt{S},$

(19.19)

which is sometimes called the *square root law* (see, e.g., BOMFORD [1971]).

Putting several levelling lines together (FIG. 4) to form a closed loop allows us to write the equation for the actual *levelling misclosure* as

$$w = \sum_{i=1}^m \Delta H_i. \quad (19.20)$$

Clearly, its expected value is zero, i.e.,  $E(w) = 0$ , and its variance is expected to be

$$\sigma_w^2 = \sigma_1^2 \sum_{i=1}^m S_i, \quad (19.21)$$

where  $\sum_{i=1}^m S_i$  is the perimeter of the loop. The standardized misclosures are

$$\tilde{w} = \frac{w}{\sigma_1 (\sum_{i=1}^m S_i)^{1/2}}. \quad (19.22)$$

These have a standard normal distribution  $n(\xi; 0, 1)$  if the  $w$ 's have a normal distribution  $n(\xi; 0, \sigma_w^2)$ .

For most of the existing national height networks, the actual standard deviation of the standardized circuit misclosures is significantly larger than 1 (cf. LUCHT [1972]). This has been a long known fact acknowledged even in the specifications for surveys of different orders (see, e.g., U.S. FEDERAL GEODETIC CONTROL COMMITTEE [1974]). Of the several explanations for this phenomenon that can be found in the literature, the following is the most likely: The basic assumption of statistical independence of individual height differences ( $\delta H$ ) between consecutive bench marks along a given levelling line is not satisfied [LUCHT, 1972; REMMER, 1975; VANÍČEK AND GRAFARENDS, 1980]. As we shall see, this explanation also covers the possibility of the presence of some unmodelled, systematic effects.

Let us write the height difference between two junction points as

$$\Delta H = \sum_{i=1}^n \delta H_i = \mathbf{u} \boldsymbol{\delta H}, \quad (19.23)$$

and its variance as

$$\sigma_{\Delta H}^2 = \mathbf{u} \mathbf{C}_{\delta H} \mathbf{u}^T, \quad (19.24)$$

where  $\mathbf{u}$  is a vector of ones, and  $\mathbf{C}_{\delta H}$  is the covariance matrix of the height differences  $\delta H$ . Now, for simplicity, assume the lengths of sections  $\Delta S_i$  to be equal. Then, in terms of correlation coefficients  $\rho_i$  ( $\rho_i = \sigma_{kl}/\sigma_k \sigma_l$  and  $i = |k - l|$ ),

$$\mathbf{C}_{\delta H} = \sigma_1^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & & & \\ \rho_2 & & 1 & & \\ \vdots & & & \ddots & \vdots \\ \rho_{n-1} & \dots & & & 1 \end{bmatrix} \quad (19.25)$$

Substitution of (25) into (24) yields

$$\sigma_{\Delta H}^2 = \sigma_1^2 \left[ n + 2 \sum_{i=1}^{n-1} (n-i)\rho_i \right]. \quad (19.26)$$

For  $\rho_i = 0$ , we get the statistically independent case (19), and for  $\rho_i = 1$ , we get the totally dependent height differences which is the other limiting case; namely,

$$\sigma_{\Delta H}^2 = \sigma_1^2 n^2 = \sigma_1^2 S^2. \quad (19.27)$$

Thus we can write the following inequality which reflects the bounds for the variance of a levelling line (also see FIG. 9):

$$\sigma_1^2 S \leq \sigma_{\Delta H}^2 \leq \sigma_1^2 S^2. \quad (19.28)$$

The remaining issue left to be discussed is the construction of the covariance matrix  $\mathbf{C}_{\delta H}$ . As explained in §10.4, this matrix is intimately related to the covariance function. One possible family of covariance functions applicable here has been postulated by LUCHT [1972] and others to be (cf. FIG. 10)

$$\text{cov}(\lambda; |S - S'|) = \lambda^{|S - S'|} = \lambda^{S_s}, \quad (19.29)$$

where  $\lambda$  is the only parameter of the function, and  $|S - S'| = S_s$  is the distance between, say, the midpoints of the two levelling segments in question. Evidently, the family of error propagation curves corresponding to the family of covariance functions has the limits  $\sqrt{S}$  and  $S$  of FIG. 9. This fact has led geodesists to declare that the *power law*,

$$\sigma_{\Delta H}/\sigma_1 = S^\alpha, \quad 0.5 \leq \alpha \leq 1, \quad (19.30)$$

governs the propagation of statistically dependent errors [MÜLLER AND SCHNEIDER, 1968].

By utilizing the above power law and the postulated family of covariance functions, it is possible to compute the elements of the  $\mathbf{C}_{\delta H}$  as follows:

(a) Given the estimate of the average value of  $\sigma_{\Delta H}/\sigma_1$  for a certain region (which is assumed to reflect the common general conditions, such as climate, in the levelled area) along with the particular value  $S$ , the quantity  $S^\alpha$  may be computed using (30), and then the variances (using (30) again) are computed from

$$\sigma_i^2 = (\sigma_1 S_i^\alpha)^2, \quad i = 1, \dots, n. \quad (19.31)$$

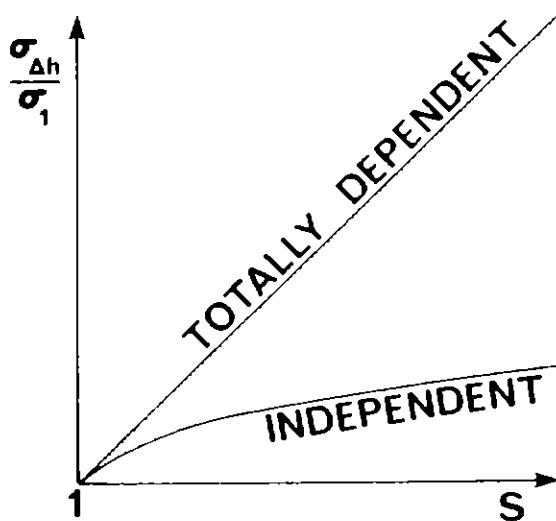


FIG. 19.9. Family of laws of propagation of errors.

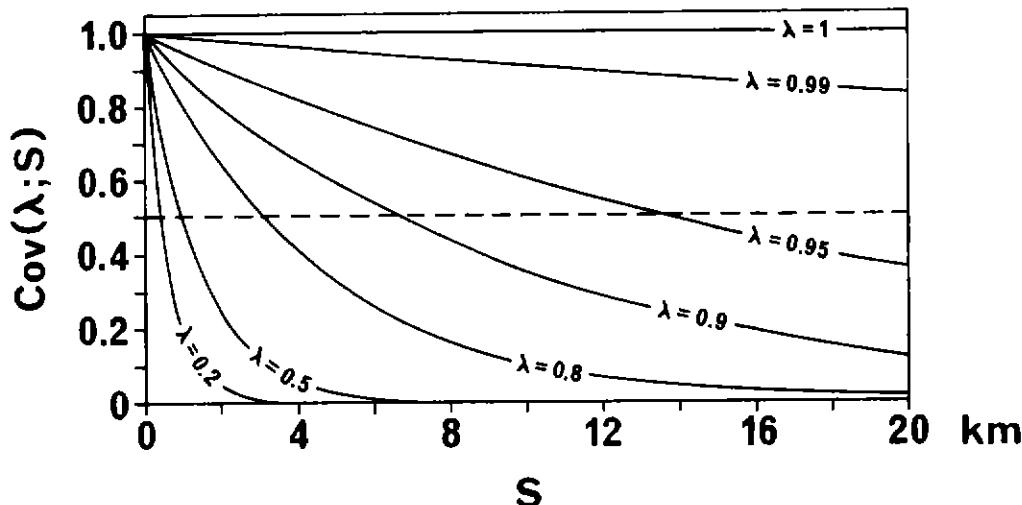


FIG. 19.10. Family of covariance functions.

(b) Again, using the general value of  $\sigma_{\Delta H}/\sigma_1$  for the region, and the particular value  $S$  for the line in question, a value of  $\lambda$  may be obtained from a nomogram (cf. FIG. 11), and the quantity

$$\text{cov}(\lambda; |S - S'|) = \lambda^{S_s}, \quad (19.32)$$

is computed using  $S_s$ . This makes it possible to determine the covariance elements

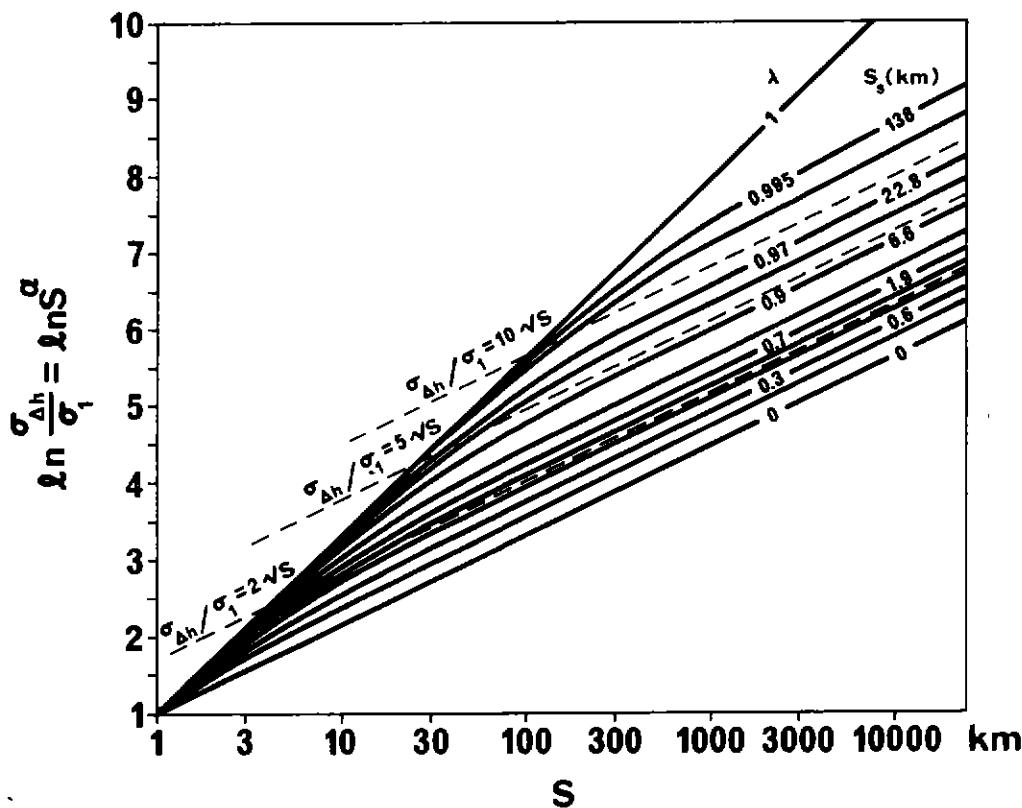


FIG. 19.11. Determination of the average parameter of statistical dependence.

from (cf. (10.43))

$$\sigma_{ij} = \sigma_i \sigma_j \text{cov}(\lambda; |S - S'|) \quad i, j = 1, \dots, n. \quad (19.33)$$

We hasten to point out that the above are only the concepts for handling problematic statistically dependent errors in levelling. Research is needed into finding the best kind of covariance function for a given region. Clearly, a similar approach should be used in obtaining the covariance matrix  $C_{\Delta H}$  corresponding to the lines between all the junction points. These fully populated covariance matrices should be used in the adjustment of the network.

Once the height differences  $\Delta H$  of the junction points have been adjusted, the observed individual height differences  $\delta H$  within each levelling line have to be corrected. In this situation, we speak of *back distribution of the residuals* from the adjustment. Since a similar problem is also encountered elsewhere in geodesy (e.g., in densification of horizontal networks by traverses or in the dynamic adjustment of height networks—see §26.4), it should pay us to have a closer look at it here.

Let us denote the adjusted heights of  $P_1$  and  $P_2$  by  $\hat{H}_1$  and  $\hat{H}_2$ . For the adjusted height  $\hat{H}_s$  of a point at a distance  $S$  from  $P_1$  (cf. FIG. 12), we can then write

$$\begin{aligned} \hat{H}_s &= \hat{H}_1 + \delta H_{1s} + \frac{S}{S_{12}} (\delta \hat{H}_{12} - \delta H_{12}) \\ &= \frac{S_{12} - S}{S_{12}} \hat{H}_1 + \frac{S}{S_{12}} \hat{H}_2 + \delta H_{1s} - \frac{S}{S_{12}} \delta H_{12}. \end{aligned} \quad (19.34)$$

Denoting  $S/S_{12}$  by  $q \in \langle 0, 1 \rangle$ , we can rewrite (34) as

$$\begin{aligned} \hat{H}_s &= [1 - q, q] [\hat{H}_1, \hat{H}_2]^T + \sum_{i=1}^{k_s} \delta H_{i,i+1} - \frac{S}{S_{12}} \sum_{i=1}^n \delta H_{i,i+1} \\ &= [1 - q, q] [\hat{H}_1, \hat{H}_2]^T + (1 - q) \sum_{i=1}^{k_s} \delta H_{i,i+1} - q \sum_{i=k_s}^n \delta H_{i,i+1}. \end{aligned} \quad (19.35)$$

Applying the covariance law (11.17) to the first term and the error propagation law (11.20) to the other two, i.e., assuming for simplicity the statistical independence of all  $\delta H$ , we get

$$\sigma_{\hat{H}_s}^2 = [1 - q, q] C_{\hat{H}_1 \hat{H}_2} [1 - q, q]^T + \hat{\sigma}_0^2(H)(1 - q)^2 S + \hat{\sigma}_0^2(H)q^2(S_{12} - S). \quad (19.36)$$

Realizing that the covariance matrix of  $\hat{H}_1, \hat{H}_2$  is a part of the covariance matrix of the network after the adjustment, namely,

$$C_{\hat{H}_1 \hat{H}_2} = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{bmatrix}, \quad (19.37)$$

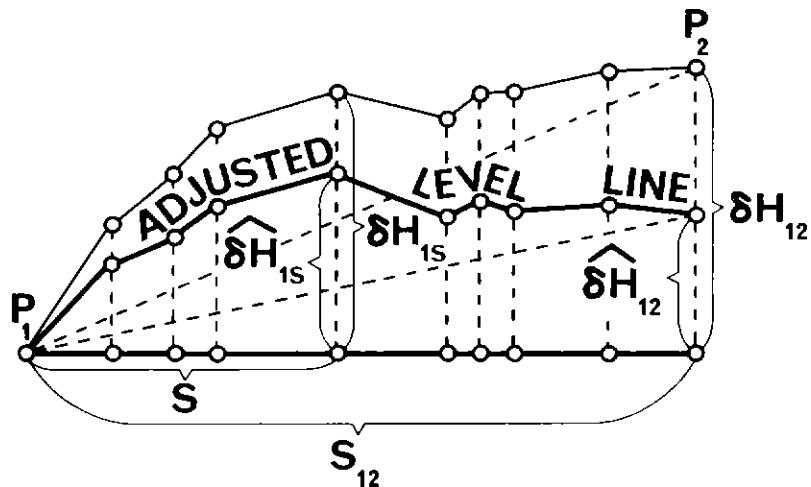


FIG. 19.12. Adjustment of a levelling line.

we finally get

$$\sigma_{\hat{H}_s}^2 = (1-q)^2 \hat{\sigma}_1^2 + 2q(1-q) \hat{\sigma}_{12} + q^2 \hat{\sigma}_2^2 + q(1-q) \hat{\sigma}_{\delta H_{12}}^2, \quad (19.38)$$

where  $\hat{\sigma}_{\delta H_{12}}^2 = S_{12} \hat{\sigma}_0^2(H)$  is the a posteriori estimate of the variance of  $\delta H_{12}$ .

It is interesting to have a look at the behaviour of  $\sigma_{\hat{H}_s}^2$  for  $S=0$  (i.e.,  $P_1$ ),  $S=\frac{1}{2}S_{12}$  (i.e., the midpoint), and  $S=S_{12}$  (i.e.,  $P_2$ ). From (38) we can write directly

$$\sigma_{\hat{H}_s}^2 = \begin{cases} \hat{\sigma}_1^2, & S=0, \\ \frac{1}{4}(\hat{\sigma}_1^2 + 2\hat{\sigma}_{12}^2 + \hat{\sigma}_2^2 + \hat{\sigma}_{\delta H_{12}}^2), & S=\frac{1}{2}S_{12}, \\ \hat{\sigma}_2^2, & S=S_{12}. \end{cases} \quad (19.39)$$

Evidently, for both junction points  $P_1, P_2$  we get the expected answers. For the midpoint, the first three terms can be shown to represent the variance of  $\frac{1}{2}(\hat{H}_1 + \hat{H}_2)$ , and the last term is the variance of  $\frac{1}{2}\delta\hat{H}_{12}$ . Thus the error at the midpoint is larger than at any of the junction points, which conforms with the findings pertaining to traverses (cf. FIG. 18.9).

### 19.3. Assessment and design of height networks

The assessment of height networks consists of the evaluation of the random and systematic effects on the adjusted heights. Beginning with the assessment of random errors, it should be clear to the reader that the situation here is analogous to that of the horizontal networks (cf. §18.3). The only difference is that here the confidence region is described by an interval while in horizontal networks it is described by an ellipse.

The *out-of-context point confidence interval* for point  $P_i$  in the network is given by

$$|H_i - \hat{H}_i| \leq C_\alpha \hat{\sigma}_{H_i}, \quad (19.40)$$

where, of course,  $\hat{H}_i$ ,  $\hat{\sigma}_{H_i}$  are the least-squares estimates of the height of  $P_i$  and its standard deviation. The expansion factor  $C_\alpha$  is given as

$$C_\alpha = \sqrt{\xi_{\chi^2_{m-1}, 1-\alpha}}, \quad (19.41)$$

if  $\sigma_0^2$  is used, or

$$C_\alpha = \sqrt{\xi_{F_{1,m-u}, 1-\alpha}}, \quad (19.42)$$

if  $\hat{\sigma}_0^2$  is used. Here  $m$  is the number of observed height differences between adjacent junction points, and  $u$  is the number of junction points. The *in-context (simultaneous) point confidence intervals* differ from the out-of-context variety only by having the probability  $\alpha$  replaced by  $\alpha/N$ .

The relative error ellipses of §18.3 have their counterparts in *relative confidence intervals*; these are simply confidence intervals for individual height differences. We get

$$|\Delta H_{ij} - \Delta \hat{H}_{ij}| \leq C_\alpha \hat{\sigma}_{\Delta H_{ij}}. \quad (19.43)$$

These again may be of the out-of-context or simultaneous varieties. As before, the neglect of covariances in the covariance matrices used leads to Bonferroni's inequality.

The *assessment of systematic network distortions* due to various causes is a much more difficult proposition. It can be done either by modelling the distortions, as discussed in §19.2, or by studying the misclosures of levelling circuits and the degree of statistical dependence of levelled height differences. One such study, conducted by REMMER [1975] should be mentioned here. He found a high degree of correlation between forward and backward runnings ( $\rho = 0.72$ ) in the Danish first-order network. This approach can, however, only detect the presence of distortions without revealing their provenance. Studies of the statistical distribution of discrepancies between forward and backward runnings detected skewness [WASSEF AND MESSIH, 1960] and bimodality [VANÍČEK AND HAMILTON, 1972] also pointing in the same direction.

Let us now turn our attention to *height network design*. It is the custom to design the networks to minimize the effect of random errors, assuming that the systematic distortions can be removed. Here we examine the error propagation for three different configurations: lines, chains of loops, and areal networks of loops. Plotted in FIG. 13 are the relative standard confidence intervals and accumulated absolute standard confidence intervals for each configuration evaluated for constant variance of observed height differences equal to one unit. Relative accuracy in terms of  $\hat{\sigma}_{\Delta H}$  ranges between 0.57 and 1 and, as expected, is best for the areal configuration. Using the *random walk theory*, BORRE AND MEISL [1974] have proved that, for an infinite areal network of regular square loops, the best relative accuracy attainable is

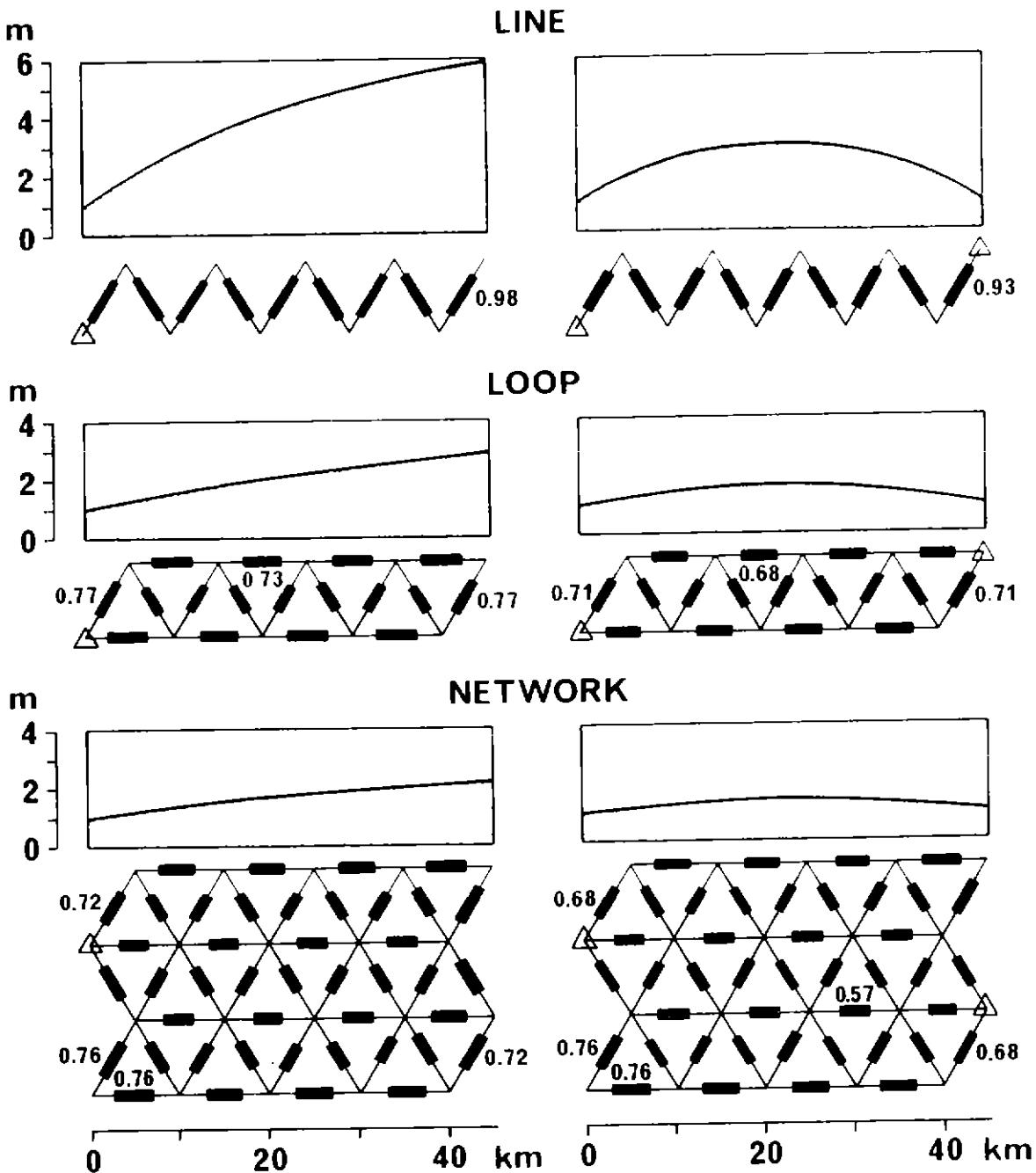


FIG. 19.13. Propagation of random errors in vertical networks. (For an explanation, see the text.)

0.5, while for triangular loops, the value is 0.33. The key to obtaining a high relative accuracy is to increase the number of lines emanating from each point. It is useful to know that a minor deviation from the above ideal geometrical forms would not appreciably change the results given above.

It is possible to strengthen a height network by including weighted constraints in the adjustment in terms of height differences of bench marks on opposite shores of a lake. These height differences can be determined through *water transfer*, a technique that uses lake water gauge records. By this technique, one can determine the

appropriate simultaneous height differences of the two shores, much the same way as one determines the local mean sea level (cf. §19.1). The concept behind this technique is that the lake level goes up and down simultaneously at all shores. The accuracy of water transfers is almost as good as that of the first-order levelling.

In particularly weak spots of the network, it may be advantageous to include heights derived from three-dimensional positions determined from terrestrial or extraterrestrial observations. The main problems with these height constraints are the necessity to know the geoidal heights, and, of course, the proper assessment of their weights vis-à-vis the weights of the levelled height differences. The methodology needed is given in §14.6.

The last topic we want to include in this section is the *densification of height networks*. The philosophy of densification, as well as the mathematical technique is the same as that of horizontal networks. Let us only state here that, in contrast to horizontal networks, there is much less to be gained by including the densification networks in the adjustment of the higher order height networks. The strengthening achieved by doing it is much less significant. Finally, it should be mentioned that most of the densification of height networks nowadays is done through aerotriangulation (see §17.2).

#### 19.4. Other heighting concepts

In this section, which is only very loosely related to the preceding three, we discuss the concepts behind the determination of heights by other methods. These remaining methods use different physical principles in determining the height of the ocean floor, sea surface, or terrain with respect to some datum surface. Let us literally begin from the bottom and work our way up by first discussing (a) bathymetry, then (b) barometric levelling, (c) airborne profiling, and lastly (d) satellite altimetry.

(a) *Bathymetry*, or depth determination, utilizes acoustic waves (see §9.2) transmitted from a surface survey vessel down to the sea floor [INGHAM, 1974]. Clearly, the datum to which the measured depth is referred is the instantaneous sea surface; as this surface is time varying, the measurements have to be reduced to a stationary surface. The two main stationary surfaces used here are the *sounding datum*, the surface to which the measurements are referred, and the *chart datum*, the surface to which the depths are reduced for the users [HYDROGRAPHER OF THE NAVY, 1965]. The main criterion for the selection of the chart datum is that, even at low tide, the actual water level should seldom fall below it. This means that the study of sea tides is of paramount importance in this regard.

The depth  $d$  at any point along the ship's track is determined from

$$d = \frac{1}{2} \int_{\tau_i}^{\tau_r} c(\tau) d\tau, \quad (19.44)$$

where  $c(\tau)$  is the actual velocity of the acoustic wave in a column of sea water, and  $\tau_i$  and  $\tau_r$  are the times of wave transmission and reception. In practice,  $c(\tau)$  is

expressed as a function of temperature  $t$ , salinity  $s$ , and hydrostatic pressure  $p$ , as [HILL, 1966]

$$c(\tau) = c_0 + \Delta c_s(\tau) + \Delta c_t(\tau) + \Delta c_p(\tau) + \Delta c_{s,t,p}(\tau). \quad (19.45)$$

Here,  $c_0$  is some standard velocity, while the remaining terms express the departures from this standard value. The main problem in evaluating the actual velocity is the availability of salinity, temperature, and pressure information. This can be obtained either from direct measurements or from models. FIG. 14 shows one idealized and two extreme examples of actual variations of acoustic velocity with depth. In the deep ocean, velocity has been tabulated as a function of position [MATTHEWS, 1939].

Bathymetric instrumentation consists of high frequency, narrow band *sonar devices*, or echo sounders, whose beam width is one of their main characteristics [MACPHEE, 1976]. Wide-beam, typically  $30^\circ$ , echo sounders are used in charting shallow waters for navigational hazards; in this application, it is the shallowest depth that is of interest. Narrow-beam,  $5^\circ$  or less, echo sounders give better defined *soundings* (depths) and are used to map the deep sea floor. At present, the highest achievable depth accuracy is about one order of magnitude better than that of marine horizontal positions (see §18.4) and by a factor of four better than routine bathymetry [COMMITTEE ON GEODESY, 1978]. THOMSON AND WELLS [1977] quote the following typical accuracy values for various depths:  $d = 10$  m,  $0.3 \text{ m} < \sigma_d < 1.0 \text{ m}$ ;  $d = 30$  m,  $0.3 \text{ m} < \sigma_d < 1.1 \text{ m}$ ; and  $d = 100$  m,  $0.6 \text{ m} < \sigma_d < 2.3 \text{ m}$ .

There is another method of determining the depth of shallow water. It involves *two-media photography* (air and water) taken from an aircraft whose position is monitored by an inertial device [REID ET AL., 1977]. The photogrammetrical process

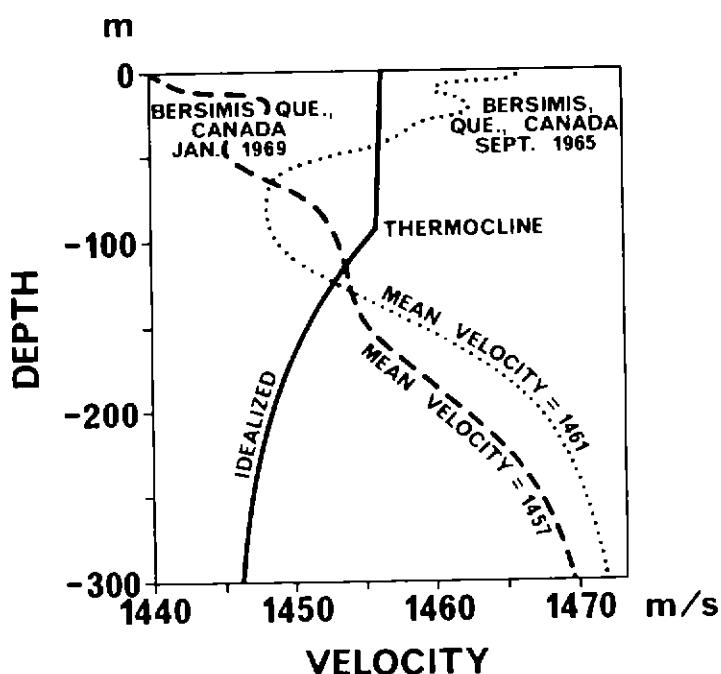


FIG. 19.14. Acoustic velocity profiles (according to THOMSON AND WELLS [1977]).

utilizes an instrument called an analytical stereoplotter which, among the other standard functions, employs an analytical correction for two-media refraction. Using this approach, the depth of coastal water can be determined to an accuracy of about  $\sigma = 1$  metre.

(b) *Barometric heighting* is based on the known physical relations among atmospheric pressure  $p$ , gravity  $g$ , air density  $\rho$ , and height  $H$  (see §9.1). Assuming the air temperature to be constant within the area of interest, we get

$$p = f(g, \rho, H). \quad (19.46)$$

In so far as the pressure variation is concerned,  $H$  is the most significant parameter in the above relationship. This fact allows us to write the mathematical model in a differential form (cf. (9.2)) as

$$dH = -\frac{dp}{gp}. \quad (19.47)$$

Utilizing the known gas laws, we can write

$$H_2 - H_1 = \int_{H_1}^{H_2} dH = \int_{p_2}^{p_1} \frac{p_s T dp}{g_s \rho_s T_s p} = \frac{H_s T}{T_s} \int_{p_2}^{p_1} \frac{dp}{p}. \quad (19.48)$$

Finally,

$$H_2 - H_1 = H_s \frac{T}{T_s} (\ln p_1 - \ln p_2), \quad (19.49)$$

where the terms subscripted with s refer to standard atmospheric conditions, and  $T$  is the actual temperature. The above model is known as *Laplace's barometric equation*, valid if the isothermal assumption is valid.

In the above development, the two points in question were assumed to lie on the same plumb line. When applying the model to two points that are not located along the same plumb line, it is necessary to further assume that the isobaric surfaces (cf. §9.1) of these two points are parallel. This assumption is usually valid within an area of a few kilometres. ALLAN ET AL. [1968] give a more complete mathematical model to allow for the spatial variation of gravity and to account for the fact that air is not a perfect gas but contains water vapour. For calibration of barometers and field procedures, the interested reader is referred to ALLAN ET AL. [1968].

When the proper measuring technique is used, the typical accuracy of barometric heighting is about  $\sigma = 2$  m in both flat and rolling terrain. In mountainous terrain, relatively large changes occur in all quantities involved, and the accuracy deteriorates accordingly. Barometric heighting, being a discrete operation, does not yield continuous height profiles as do the aforementioned bathymetric techniques.

(c) *Airborne profile recording* (APR), as the name suggests, is a technique whereby a continuous profile of the terrain below a flying aircraft is recorded. The general operation is described in THOMPSON [1966] as follows: The aircraft first climbs to a

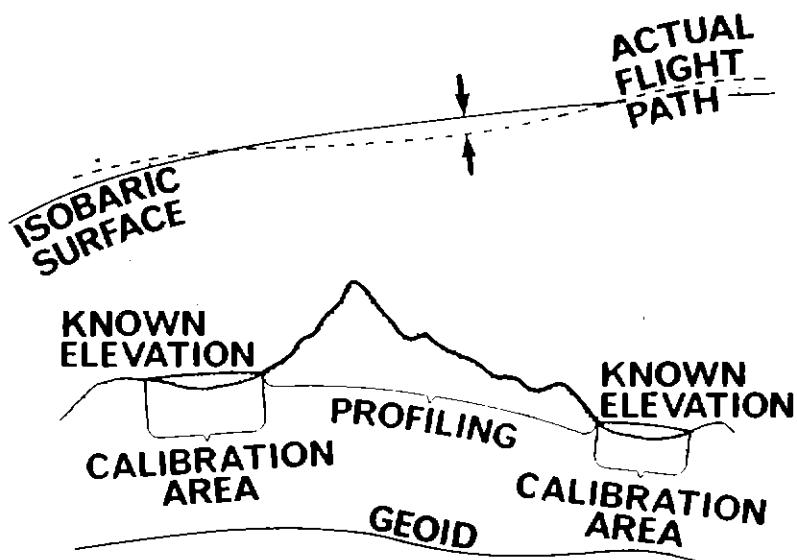


FIG. 19.15. Airborne profiling.

height of some 3000 m above the terrain. There the system is calibrated over an area of known heights, usually a lake surface (see FIG. 15), and the differential barometer is set to refer to an appropriate reference pressure. The aircraft is then flown as close as possible (practically to within 20 m) to the surface of the selected reference pressure, i.e., to the reference isobaric surface, and the terrain profile is recorded with an airborne profile recorder. At the same time, the variations of pressure, which can be translated into the deviations of the flight path from the isobaric surface, are also recorded. At the end of the run, the system is calibrated again.

Present airborne profile recorders use pulsed microwaves and have an accuracy of  $\sigma = 3$  metres. This accuracy is limited by the relatively large angular dispersion of the transmitted beam, which is of the order of  $1.5^\circ$ . It causes the received reflection to have come from an area about the size of a hectare. We speak of this as a large *footprint* of the system. Microwave beams, by their nature, cannot be focussed any better, and lasers are being experimented with to solve this problem. The aim is to restrict the footprint on the ground to only a few square centimetres to ensure an accuracy of  $\sigma = 0.5$  m or better.

The main feature limiting the accuracy of the APR operation is the accuracy with which the isobaric surface can be located at any point. This can be done to about  $\sigma = 5$  m when the meteorological conditions are favourable and the isobaric surface is flat. With the laser APR systems now being developed [HURSH ET AL., 1977], the possibility of positioning the aircraft with an inertial device, or even having the aircraft tracked by satellites, is being studied; aircraft can also be positioned through the use of air photograph strips [WISE, 1979]. Such positioning of the aircraft would solve the accuracy problem arising from the use of isobaric surfaces.

(d) *Satellite altimetry* is a technique whereby a satellite orbiting above the earth both emits short bursts of electromagnetic waves downward and records the times of

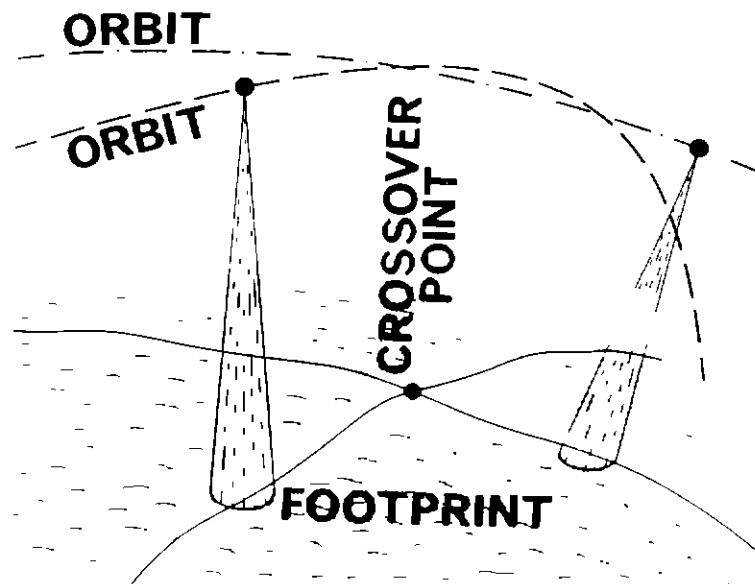


FIG. 19.16. Satellite altimetry

arrival of pulses reflected from the surface of the earth. From these timings, the vertical distances between the satellite and the earth's surface (usually the sea surface) is derived—see FIG. 16. In the meantime, the satellite is tracked from tracking stations of known positions to determine its orbit (see §23.2).

Once the height  $h_s$  of the satellite orbit above the reference ellipsoid and the satellite altitude  $a$  above the sea surface are known, they can be used to determine the height  $h_i$  of the instantaneous sea surface above the reference ellipsoid being used, i.e.,

$$h_i = h_s - a \quad (19.50)$$

(cf. FIG. 17). To this end, it is necessary to correct for orbital biases, anomalous sea state due to meteorological conditions, etc. [RUMMEL AND RAPP, 1977]. The simplest way of determining the mean sea surface is probably through the use of a network of *crossover points*, i.e., points where individual orbits pass over each other, where the corrected sea surface topography at the two epochs is required to coincide [MATHER ET AL., 1979].

Clearly, if the geoidal height  $N$  is also known, then one can obtain the instantaneous sea surface height (topography)  $H_i$  above the geoid (cf. §19.1) simply by subtracting  $N$  from  $h_i$ . Conversely, if the instantaneous sea surface topography is neglected, then  $h_i$  can be viewed as a first approximation to the geoidal height. For this application, it is advisable to correct  $h_i$  at least for the effect of sea tide. If the values of instantaneous sea surface topography, corrected for sea tide, are determined over and over again during a long period of time, then the mean sea surface topography can be derived.

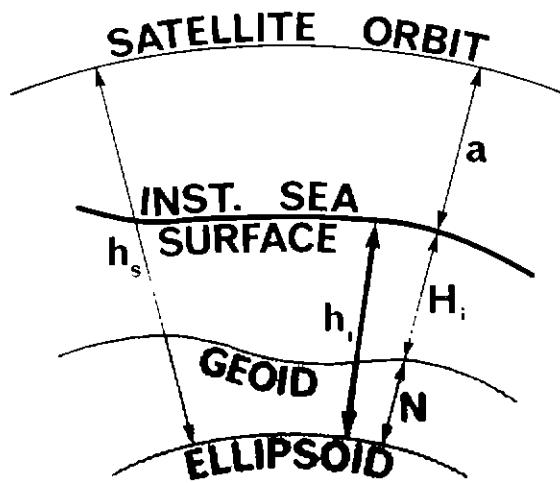


FIG. 19.17. Satellite altimetry model.

The first altimeter was flown in the SKYLAB and had an accuracy of about 10 m [McGOOGAN ET AL., 1974]. The GEOS-3 satellite altimeter was accurate to about  $\sigma = 1$  metre [STANLEY, 1979; AGU, 1979], and SEASAT [AGU, 1982; 1983] about one order of magnitude better. Recent results suggest that a longer term sea surface topography accurate to about  $\sigma = 15$  cm can be obtained from satellite altimetry. FIG. 18 shows a global compilation of results from SEASAT [CHENEY AND MARSH, 1982]. The reader is advised to compare these results with FIG. 7.8.

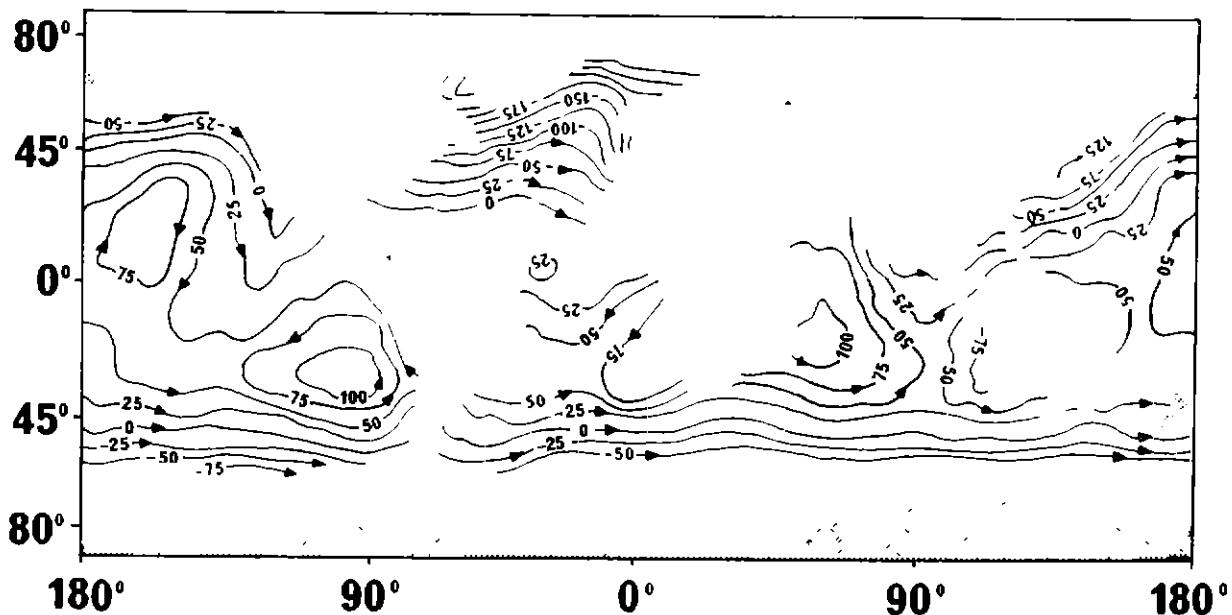


FIG. 19.18. World wide sea surface topography from SEASAT altimetry. Contours in centimetres.

## PART IV

### REFERENCES

- AARDOOM, L., A.G. GIRNIUS AND G. VEIS (1967). Determination of the absolute space directions between Baker-Nunn camera stations. *Proc. 2nd International Symposium on the Use of Artificial Satellites for Geodesy*, Ed. G. Veis, IAG and COSPAR, Athens, Greece, April, 1965. National Technical University, Vol. II, pp. 315-344.
- ACKERMANN, F. (1968). Gesetzmässigkeiten der absoluten Lagegenauigkeit von Blöcken. *Bildmessung und Luftbildwesen* 36 (1), pp. 3-15.
- ACKERMANN, F. (1974). Results of recent experimental investigations in aerial triangulation. *Proc. 40th Annual Meeting of the American Society of Photogrammetry*, St. Louis, U.S.A., March, pp. 216-234.
- ADAMS, G.W. (1977). Inertial survey data reduction using maximum likelihood estimation. *Proc. 1st International Symposium on Inertial Technology for Surveying and Geodesy*, IAG and CIS, Ottawa, Canada, October. Canadian Institute of Surveying, pp. 380-387.
- ALBERDA, J.E. (1974). Aspects of large leveling nets. *Canad. Surv.* 28 (5), pp. 643-652.
- ALLAN, A.L., J.R. HOLLOWAY AND J.H.B. MAYNES (1968). *Practical Field Surveying and Computations*. Heinemann.
- AMERICAN GEOPHYSICAL UNION (1979). *Journal of Geophysical Research* 84(B8), July, pp. 3779-4082.
- AMERICAN GEOPHYSICAL UNION (1982). *SEASAT Special Issue I*. Reprinted from *Journal of Geophysical Research* 87(C5), April.
- AMERICAN GEOPHYSICAL UNION (1983). *SEASAT Special Issue II*. Reprinted from *Journal of Geophysical Research* 88(C3), February, pp. 1529-1952.
- ANDERLE, R.J. (1974). Transformation of terrestrial survey data to Doppler satellite datum. *J. Geophys. Res.* 79 (35), pp. 5319-5331.
- ANDERLE, R.J. (1980). The global positioning system. *Philos. Trans. Roy. Soc. London Ser. A* 294, pp. 395-406.
- ANDERSON, E.G. (1978). Modelling of physical influences in sea level records for vertical crustal movement detection. *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference, An International Symposium on the Applications of Geodesy to Geodynamics*, Ed. I.I. Mueller. IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280. The Ohio State University, Columbus, U.S.A., pp. 145-152.
- ANDERSON, E.N. (1966). *Principles of Navigation*. Hollis and Carter.
- ANGUS-LEPPAN, P.V. (1972). Adjustment of trilateration using length ratios. *Surv. Rev.* XXI (166), pp. 355-368.
- Apparent Places of Fundamental Stars, 1979 (1977). Astronomisches Rechen-Institut, Heidelberg, Germany.
- ASHKENAZI, V. AND P.A. CROSS (1972). Strength analysis of block VI of the European triangulation. *Bull. Géod.* 103, pp. 5-25.
- BAARDA, W. (1973). S-transformations and criterion matrices. Netherlands Geodetic Commission, Publications on Geodesy, New Series 5 (1), Delft, Netherlands.
- BAESCHLIN, C.F. (1960). Das Geopotential, metrische Höhen und Gebrauchshöhen. *Schweizerische Zeitschrift für Vermessung, Kulturtechnik und Photogrammetrie* 58 (6).
- BARTELME, N. AND P. MEISSL (1974). Strength analysis of distance networks. Geodetic Institute of the Technical University Report 15, Graz, Austria.
- BEATTIE, D.S. (1978). Documentation of program GANET (geodetic adjustment of networks). Publication of the Geodetic Survey of Canada, Department of Energy, Mines and Resources, Ottawa, Canada.

- BJERKNES, V. AND J.W. SANDSTRÖM (1910). Dynamical meteorology and hydrography. Part I, statistics. Publications of the Carnegie Institute 88, pp. 1-146, Washington, D.C., U.S.A.
- BLAHA, G. (1971a). Inner adjustment constraints with emphasis on range observations. Department of Geodetic Science Report 148, The Ohio State University, Columbus, U.S.A.
- BLAHA, G. (1971b). Investigations of critical configurations for fundamental range networks. Department of Geodetic Science Report 150, The Ohio State University, Columbus, U.S.A.
- BLAIS, J.A.R. (1979). Least-squares block adjustment of stereoscopic models and error analysis. Ph.D. dissertation, Department of Surveying Engineering, University of New Brunswick, Fredericton, Canada.
- BOMFORD, G. (1971). *Geodesy*. 3rd ed., Oxford University Press.
- BORRE, K. (1977). Geodetic elasticity theory: its matter and an application. *Bull. Géod.* 51 (1), pp. 63-71.
- BORRE, K. (1978). Error propagation in absolute geodetic networks—a continuous approach. *Studia Geoph. et Geod.* 22, pp. 213-223.
- BORRE, K. AND P. MEISSL (1974). Strength analysis of leveling-type networks. An application of random walk theory. Danish Geodetic Institute Report 50, Copenhagen, Denmark.
- BOSSLER, J.D., C.C. GOAD AND P.L. BENDER (1980). Using the Global Positioning System for geodetic positioning. *Bull. Géod.* 54, pp. 553-563.
- BOWDITCH, N. (1977). *American Practical Navigator: An Epitome of Navigation*. Defense Mapping Agency Hydrographic Center Publication 9, DMA stock no. NVPUB9V1, Washington, D.C., U.S.A.
- BREMNER, H. (1949). *Terrestrial Radio Waves—Theory of Propagation*. Elsevier.
- BRITTING, K.R. (1971). *Inertial Navigation Systems Analysis*. Wiley.
- BRODEN, N.W., T.H. LEGG AND J.L. LOCKE (AND OTHERS) (1967). Long base line interferometry: a new technique. *Science* 156 (3782), pp. 1592-1593.
- BROUWER, D. AND G.M. CLEMENCE (1961). *Methods of Celestial Mechanics*. Academic Press.
- BROWN, D.C. (1970). Near term prospects for positional accuracies of 0.1 to 1.0 metres from satellite geodesy. Report prepared by DBA Systems, Inc. for Air Force Cambridge Research Laboratories, Report AFCRL-70-0501, Bedford, U.S.A.
- BROWN, D.C. AND J.E. TROTTER (1969). SAGA, a computer program for short arc geodetic adjustment of satellite observations. Prepared by DBA Systems, Inc. for Air Force Cambridge Research Laboratories, Report AFCRL-69-0080, Bedford, U.S.A.
- BRUNAVS, P. AND D.E. WELLS (1971). Accurate phase lag measurements over seawater using Decca Lambda. Unpublished manuscript, Atlantic Oceanographic Laboratory, Bedford Institute, Dartmouth, Canada.
- BRUNNER, F.K. (1980). Systematic and random atmospheric refraction effects in geodetic levelling. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Vertical Geodetic Networks*, Ed. G. Lachapelle, Canadian Department of Energy, Mines and Resources, CIS, NSERC, Ottawa, Canada, May, pp. 691-704.
- BRUNS, H. (1878). Die Figur der Erde. Publication des Königlichen Preussischen Geodätischen Institutes, Berlin, Germany.
- BURSA, M. (1962). The theory of the determination of the non-parallelism of the minor axis of the reference ellipsoid, polar axis of inertia of the earth, and initial astronomical and geodetic meridians from observations of artificial earth satellites. Translated from Russian by the National Translation Center, The John Crerar Library, Chicago, U.S.A., from *Stud. Geoph. et Geod.* 6, pp. 209-214.
- BURŠA, M. (1965). Determination of the direction of the minor axis of the reference ellipsoid and the plane of the initial geodetic meridian from the artificial satellite data. 1973 translation from Russian by the National Research Council of Canada, from *Stud. Geoph. et Geod.* 9 (1), pp. 14-22.
- CANNON, J.B. (1929). Adjustment of the precise level net of Canada 1928. Geodetic Survey of Canada Special Publication 28, Department of Energy, Mines and Resources, Ottawa, Canada.
- CASTLE, R.O. AND P. VANIČEK (1980). Interdisciplinary considerations in the formulation of the new North American vertical datum. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Vertical Geodetic Networks*, Ed. G. Lachapelle, Canadian Department of Energy, Mines and Resources, CIS, NSERC, Ottawa, Canada, May, pp. 285-300.
- CHENEY, R.E. AND J.G. MARSH (1982). Global ocean circulation from satellite altimetry. *EOS Trans. AGU* 63, p. 997.

- CHOVITZ, B. (1974). Three-dimensional model based on Hotine's *Mathematical Geodesy*. *Canad. Surv.* 28 (5), pp. 568-573.
- CHRISTODOULIDIS, D.C. AND D.E. SMITH (1983). The role of satellite laser ranging through the 1990's. *Proc. IAG Symposia*, IAG, IUGG, Hamburg, FRG, August, Dept. of Geodetic Science and Surveying, The Ohio State University, Columbus, U.S.A., Vol. 2, pp. 408-431.
- CHRZANOWSKI, A.J. AND G. KONECNY (1965). Theoretical comparison of triangulation, trilateration and traversing. *Canad. Surv.* XIX (4), pp. 353-366.
- CLARK, D. (1969). *Plane and Geodetic Surveying for Engineers*. Vol. II, 6th ed., Constable.
- COMMITTEE ON GEODESY (1978). Geodesy: trends and prospects. U.S. National Research Council, Washington, D.C., U.S.A.
- CONTE, S.D. AND C. DE BOOR (1972). *Elementary Numerical Analysis*. McGraw-Hill.
- COUNSELMAN, C.C. AND S.A. GOUREVITCH (1981). Miniature interferometer terminals for Earth surveying: Ambiguity and multipath with Global Positioning System. *IEEE Trans. on Geoscience and Remote Sensing* GE-19(4), October.
- CURRIE, R.G. (1975). Period,  $Q_p$  and amplitude of the pole tide. *Geophys. J. Roy. Astronom. Soc.* 43, pp. 73-86.
- DARE, P. AND P. VANIČEK (1982). Strength analysis of horizontal networks using strain. *Survey Control Networks*, Proc. of meeting of Study Group 5B, Ed. K. Borre, W.M. Welsch, FIG, Aalborg University Centre, Denmark, July. Hochschule der Bundeswehr, München, Heft 7, pp. 181-196.
- DEPARTMENT OF ENERGY, MINES AND RESOURCES (1973). Specifications and recommendations for control surveys and survey markers. Surveys and Mapping Branch Misc. Ser. 73/3, Ottawa, Canada
- DEPARTMENT OF ENERGY, MINES AND RESOURCES (1979). Personal communication. Geodetic Survey of Canada, Ottawa, Canada
- DEPARTMENT OF MINES AND TECHNICAL SURVEYS (1955). Geodetic application of Shoran. Geodetic Survey of Canada Publication 78, Ottawa, Canada.
- DRACUP, J.F. (1978). Net adjustment of the NAD and the surveyor. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*. U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 481-486.
- DRAPER, C.S. (1977). Inertial technology for surveying and geodesy. *Proc. 1st International Symposium on Inertial Technology for Surveying and Geodesy*, IAG and CIS, Ottawa, Canada, October. Canadian Institute of Surveying, pp. 5-41.
- DUFOUR, H.M. (1970). Générations et applications des tableaux de variance des systèmes de moindres carrés. *Bull. Géod.* 98, pp. 309-339.
- EATON, R.M., D.E. WELLS AND N. STUIFBERGEN (1976). Satellite navigation hydrography. *Internat. Hydrogr. Rev.* LIII (1), pp. 99-116.
- EBNER, H. (1975). Selfcalibrating block adjustment by independent models. *Proc. 41st Annual Meeting of the American Society of Photogrammetry*, Washington, D.C., U.S.A., March, pp. 30-38.
- ENVIRONMENT CANADA (1979). Personal communication. Marine Environmental Data Service Branch, Marine Information Directorate, Ocean and Aquatic Sciences, Ottawa, Canada.
- FILA, K. AND C. CHAMBERLAIN (1978). Integration of secondary networks in the Maritime Provinces. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 519-527.
- FUBARA, D.M.J. (1972). Three-dimensional geodesy for terrestrial network adjustment. *J. Geophys. Res.* 77 (5), pp. 796-807.
- GARFINKEL, B. (1944). An investigation in the theory of astronomical refraction. *Astronom. J.* 50 (8).
- GOAD, C.C. AND B.W. REMONDI (1984). Initial relative positioning results using the Global Positioning System. *Bull. Géod.* 58, pp. 193-210.
- GREGERSON, L.F. (1975). Inertial geodesy in Canada. Paper presented at the Fall Meeting of the American Geophysical Union, San Francisco, U.S.A., December.
- GREGERSON, L.F. (1980). Personal communication.
- HALMOS, F. AND I. KÁDÁR (1977). An attempt to interpret physically the notion-system of geodetic information. *Bull. Géod.* 51 (1), pp. 1-16.

- HARMAN, H.H. (1967). *Modern Factor Analysis*. 2nd ed. rev., University of Chicago Press.
- HEISKANEN, W.A. AND H. MORITZ (1967). *Physical Geodesy*. Freeman.
- HELA, I. AND E. LISITZIN (1967). A world mean sea level and marine geodesy. *Proc. 1st Marine Geodesy Symposium*, Battelle Memorial Institute, Columbus, U.S.A., September, 1966. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 71–73.
- HELMERT, F.R. (1880). *Die mathematischen und physikalischen Theorien der höheren Geodäsie*. Vol. I, Minerva G.M.B.H. reprint, 1962.
- HILL, M.N. (Ed.) (1966). *The Sea*. Vol. I, Wiley Interscience.
- HOLDAHL, S.R. (1980). A model of temperature stratification for correction of levelling refraction. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Vertical Geodetic Networks*, Ed. G. Lachapelle, Canadian Department of Energy, Mines and Resources, CIS, NSERC, Ottawa, Canada, May, pp. 647–676.
- HOPFIELD, H.A. (1969). Two-quartic tropospheric refractivity profile for correcting satellite data. *J. Geophys. Res.* 74, pp. 4487–4499.
- HOSKINSON, A.J. AND J.A. DUERKSEN (1952). Manual of geodetic astronomy-determination of longitude, latitude, and azimuth. U.S. Coast and Geodetic Survey Special Publication 237, Washington, D.C., U.S.A.
- HOTHEM, L.D., D.S. ROBERTSON AND W.E. STRANGE (1978). Orientation and scale of satellite Doppler results based on combination and comparison with other space systems. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April, U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 167–180.
- HOTINE, M. (1946). The orthomorphic projection of the spheroid. *Emp. Surv. Rev.* 8, pp. 300–311.
- HOTINE, M. (1947). The orthomorphic projection of the spheroid. *Emp. Surv. Rev.* 9, pp. 25–35, 52–70, 112–123, 157–166.
- HOTINE, M. (1969). *Mathematical Geodesy*. ESSA Monograph 2. U.S. Department of Commerce, Government Printing Office, Washington, D.C., U.S.A.
- HRADILEK, L. (1972). Refraction in trigonometric and three-dimensional terrestrial networks. *Canad. Surv.* 26 (1), pp. 59–70.
- HUGGETT, G.R. AND L.E. SLATER (1978). Recent advances in multiwavelength distance measurement. *Proc. International Symposium on Electromagnetic Distance Measurement and the Influence of Atmospheric Refraction*, Ed. P. Richardus. IAG, Wageningen, The Netherlands, May, 1977. Rijkscommissie voor Geodesie, Delft, The Netherlands, pp. 141–152.
- HURSH, J.W., G. MAMON AND J.A. SOLTZ (1977). Aerial profiling of terrain. *Proc. 1st International Symposium on Inertial Technology for Surveying and Geodesy*, IAG and CIS, Ottawa, Canada, October. Canadian Institute of Surveying, pp. 121–130.
- HYDROGRAPHER OF THE NAVY (1965). Admiralty manual of hydrographic surveying. Vol. I, II, Royal Navy, London, U.K.
- INGHAM, A.E. (1974). *Hydrography for the Surveyor and Engineer*. Granada.
- ISNER, J.F. (1978). Helmert block initial level system. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 405–416.
- ISNER, J.F. AND G.M. YOUNG (1978). Horizontal data entry. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 233–246.
- JOHLER, J.R., W.G. KELLAR AND L.C. WALTERS (1956). Phase of the low radio frequency ground wave. (U.S.) National Bureau of Standards Circular 573.
- JONES, H.E. (1973). Geodetic datums in Canada. *Canad. Surv.* 27 (3), pp. 195–207.
- JORDAN, W. AND O. EGGERT (1962). *Handbuch der Vermessungskunde*. Bd. III. Translated from German by the U.S. Army Map Service, Washington, D.C., U.S.A.
- JORGENSEN, P.S. (1980). NAVSTAR/Global Positioning System 18-satellite constellation. *Navigation*, Journal of the (U.S.) Institute of Navigation, 27(2), pp. 89–100.

- KAULA, W. (1966). *Theory of Satellite Geodesy: Applications of Satellites to Geodesy*. Blaisdell.
- KAYTON, M. (1960). Coordinate frames in inertial navigation. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, U.S.A.
- KELLER, M. (1967). Block adjustment operation at C&GS. *Photogr. Engng.* XXXIII (11), pp. 1266–1275.
- KNIGHT, W. AND M.P. MEPHAM (1978). Report on computer programs for solving large systems of normal equations. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 357–363.
- KOLACZEK, B. AND G. WEIFFENBACH (Eds.) (1975). *Proceedings of Colloquium No. 26 on Reference Coordinate Systems for Earthdynamics*. IAU, Toruń, Poland, August, 1974. Polish Academy of Sciences, Toruń, Poland.
- KORN, G.A. AND T.M. KORN (1968). *Mathematical Handbook for Scientists and Engineers* 2nd ed., McGraw-Hill.
- KOUBA, J. (1976). Doppler leveling. *Canad. Surv.* 30 (1), pp. 21–32.
- KOUBA, J. (1980). Geodetic satellite Doppler positioning and application to Canadian test adjustment. *Philos. Trans. Roy. Soc. London. Ser. A* 294, pp. 271–276.
- KOUBA, J. AND D.E. WELLS (1976). Semidynamical Doppler satellite positioning. *Bull. Geod.* 50 (1), pp. 27–39.
- KRAKIWSKY, E.J. AND I.I. MUELLER (1966). Proposed establishment of a first-order height system in the U.S.A. Paper presented at the 47th Annual Meeting of the American Geophysical Union, Washington, D.C., U.S.A., April.
- KRAKIWSKY, E.J. AND D.B. THOMSON (1974). Mathematical models for the combination of terrestrial and satellite networks. *Canad. Surv.* 28 (5), pp. 606–615.
- KRAKIWSKY, E.J., D.E. WELLS AND B.P. KIRKHAM (1972). Geodetic control from Doppler satellite observations *Canad. Surv.* 26 (2), pp. 146–162.
- KUKKAMÄKI, T.J. (1938). Über die Nivellitische Refraktion. Finnish Geodetic Institute Publication 25, Helsinki, Finland.
- LAMBECK, K. (1971). The relation of some geodetic datums to a global geocentric reference system. *Bull. Géod.* 99, pp. 37–53.
- LANGLEY, R.B., G. BEUTLER, D. DELIKARAOGLOU, B. NICKERSON, R. SANTERRE, P. VANÍČEK AND D.E. WELLS (1984). Studies in the application of the Global Positioning System to differential positioning. Department of Surveying Engineering Technical Report 108. University of New Brunswick, Fredericton, Canada.
- LATIMER, J.H. AND E.M. GAPOSCHKIN (1977). Scalar translocation using laser range data. Paper presented at the Spring Annual Meeting of the American Geophysical Union, Washington, D.C., U.S.A., June.
- LEE, L.P. (1976). *Conformal Projections Based on Elliptic Functions*. Cartographica Monograph 16, B.V. Gutsell, Toronto, Canada.
- LEVALLOIS, J.J. (1964). Sur la fréquence des mesures de pesanteur dans les nivellements. *Bull. Géod.* 74, pp. 317–325.
- LUCHT, H. (1972). Korrelation im Präzisionsnivelllement. Wissenschaftliche Arbeiten der Lehrstuhle für Geodäsie, Photogrammetrie und Kartographie an der Technischen Universität Nr. 48, Hannover, Germany.
- MACDORAN, P.F., A.E. NIELL, K.M. ONG AND G.M. RESCH (1978). Radio interferometric geodetic networks *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 149–165.
- MACPHEE, S.B. (1976). Acoustics and echo sounding instrumentation. Canadian Hydrographic Service Technical Report 1976-1, Department of Fisheries and Oceans, Ottawa, Canada.
- MALING, D.H. (1973). *Coordinate Systems and Map Projections*. George Philip and Son Ltd.
- MATHER, R.S. (1970). The geocentric orientation vector for the Australian geodetic datum. *Geophys. J. Roy. Astronom. Soc.* 22, pp. 55–81.
- MATHER, R.S., C. RIZOS AND R. COLEMAN (1979). Remote sensing of surface ocean circulation with satellite altimetry. *Science* 205, pp. 11–17.

- MATTHEWS, D.J. (1939). Tables of the velocity of sound in pure water and sea water for use in echo-sounding and sound-ranging. Hydrographic Department of the Admiralty Manual HD-282, His Majesty's Stationery Office, London, U.K.
- MCGOOGAN, J.T., C.D. LEITAO, L.S. MILLER AND W.T. WELLS (1974). SKYLAB S-193 altimeter experiment performance, results and applications. *Proc. International Symposium on Applications of Marine Geodesy*, Battelle Memorial Institute, DMA, NASA, NOAA, NSF, ONR, Columbus, U.S.A., June. Marine Technology Society, Washington, D.C., U.S.A., pp. 291-300.
- MCLELLAN, C.D., A.E. PETERSON AND G. KATINAS (1970). GALS: geographic adjustment by least squares. A computer program to adjust horizontal control surveys. Report of the Geodetic Survey of Canada, Department of Energy, Mines and Resources, Ottawa, Canada.
- MEADE, R.H. AND K.O. EMERY (1971). Sea level as affected by river runoff, eastern U.S. *Science* 173 (3995), pp. 425-428.
- MEISSL, P. (1974). The strength of continental terrestrial networks. *Canad. Surv.* 28 (5), pp. 582-589.
- MEISSL, P. (1978). A priori prediction of roundoff error accumulation during the adjustment of the United States ground control network by the Helmert blocking technique. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 333-345.
- MENZEL, D.H. (1955). *Fundamental Formulas of Physics*. Vol. 2, Dover reprint, 1960.
- MERRY, C.L. AND P. VANÍČEK (1983). Investigation of local variations of sea-surface topography. *Marine Geodesy* 7(1-4), pp. 101-126.
- MILLER, A.R. (1958). The effects of winds on water levels on the New England coast. *Limnology and Oceanogr.* 3 (1), pp. 1-14.
- MOFFETT, J.B. (1971). Program requirements for two minute integrated Doppler satellite navigation solution. Applied Physics Laboratory Technical Memorandum TG-819-1, The Johns Hopkins University, Silver Spring, U.S.A.
- MOLODENSKIJ, M.S., V.F. EREMEEV AND M.I. YURKINA (1960). *Methods for study of the external gravitational field and figure of the earth*. Translated from Russian by the Israel Program for Scientific Translations for the Office of Technical Services, U.S. Department of Commerce, Washington, D.C., U.S.A., 1962.
- MONTGOMERY, R.B. (1937-38). Fluctuations in monthly sea level on eastern U.S. coast as related to dynamics of western North Atlantic ocean. *J. Mar. Res.* 1 (2), pp. 165-185.
- MUELLER, I.I. (1964). *Introduction to Satellite Geodesy*. Ungar.
- MUELLER, I.I. (1969). *Spherical and Practical Astronomy as Applied to Geodesy*. Ungar.
- MUELLER, I.I. (1974). Global satellite triangulation and trilateration results. *J. Geophys. Res.* 79 (35), pp. 5333-5347.
- MUELLER, I.I. AND M. KUMAR (1975). The OSU 275 system of satellite tracking station coordinates. Department of Geodetic Science Report 228. The Ohio State University, Columbus, U.S.A.
- MÜLLER, K. AND E. SCHNEIDER (1968). Nivellementswidersprüche in Statistischer Sicht. *Z. Vermessungstechnik* 16, pp. 8-15.
- NASSAR, M.M. (1977). Gravity field and levelled heights in Canada. Department of Surveying Engineering Technical Report 41, University of New Brunswick, Fredericton, Canada.
- NASSAR, M.M. AND P. VANÍČEK (1975). Levelling and gravity. Department of Surveying Engineering Technical Report 33, University of New Brunswick, Fredericton, Canada.
- NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (1984). Geodynamics. *Proc. of a Workshop*, Ed. L.S. Walter. NASA, Airlie, VA, U.S.A., February, 1983. NASA Conference Publication 2325.
- NEWCOMB, S. (1906). *A Compendium of Spherical Astronomy*. Dover reprint, 1960.
- PAUL, M.K. (1973). A note on the computation of geodetic (Cartesian) coordinates. *Bull. Géod.* 108, p. 135.
- PICK, M., J. PICHA AND V. VYSKOČIL (1973). *Theory of the Earth's Gravity Field*. Elsevier.
- RAINSFORD, H.F. (1955). Long geodesics on the ellipsoid. *Bull. Géod.* 37, pp. 12-22.
- RAMSAYER, K. (1971). Untersuchung der Genauigkeit eines Raumpolygonzugs. *Z. Vermessungswesen* 96 (10), pp. 429-439.
- RAPPLEYE, H.S. (1948). Manual of geodetic leveling. U.S. Coast and Geodetic Survey Special Publication 239, Washington, D.C., U.S.A.

- REID, D.B., S.E. MASRY AND J.R. GIBSON (1977). An inertially aided photobathymetry system. *Proc. 1st International Symposium on Inertial Technology for Surveying and Geodesy*, IAG and CIS, Ottawa, Canada, October. Canadian Institute of Surveying, pp. 361-369.
- REMMER, O. (1975). Levelling errors in Statu Nascendi. Geodaetisk Institut Report 51. Copenhagen, Denmark.
- RICHARDUS, P. AND R.K. ADLER (1972). *Map Projections for Geodesists, Cartographers and Geographers*. North-Holland.
- ROBBINS, A.R. (1962). Long lines on the spheroid. *Emp. Surv. Rev.* 16 (125), pp. 301-309.
- ROBBINS, A.R. (1976). Military engineering: field and geodetic astronomy. Vol. 13, Part 9, Ministry of Defence Army Code No. 71091, School of Military Survey, Hermitage, Newbury, Berkshire, U.K.
- RODEN, G.I. (1966). Low frequency sea level oscillations along the pacific coast of North America. *J. Geophys. Res.* 71 (9), pp. 4755-4776.
- ROSSITER, J.R. (1966). Long-term variations in sea level. In: *The Sea*, Vol. 1, Ed. M.N. Hill, Wiley Interscience.
- RUMMEL, R. AND R.H. RAPP (1977). Undulation and anomaly estimation using GEOS-3 altimeter data without precise satellite orbits. *Bull. Géod.* 51 (1), pp. 73-88.
- SAASTAMOINEN, J.J. (1967). *Electromagnetic Distance Measurement*. Hilger and Watts.
- SAASTAMOINEN, J.J. (1973). Contributions to the theory of atmospheric refraction. *Bull. Géod.* 107, pp. 13-34.
- SCARBOROUGH, J.B. (1958). *The Gyroscope: Theory and Applications*. Interscience.
- SCHMID, H.H. (1974). Worldwide geometric satellite triangulation. *J. Geophys. Res.* 79 (35), pp. 5349-5376.
- SCHUT, G.H. (1968). Review of strip and block adjustment during the period 1964-1967. *Photogr. Engrg.* XXXIV (4), pp. 344-355.
- SCHWARZ, C.R. (1969). The use of short arc orbital constraints in the adjustment of geodetic satellite data. Department of Geodetic Science Report 118, The Ohio State University, Columbus, U.S.A.
- SCHWARZ, C.R. (1978). TRAV10: horizontal network adjustment program. NOAA Technical Memorandum NOS NGS-12, National Geodetic Survey, Rockville, U.S.A.
- SIMONSEN, O. (1963). Report for the period September 1960-July 1963 on REUN. Danish Geodetic Institute Report, Copenhagen, Denmark.
- SMART, W.M. (1962). *Spherical Astronomy*. 5th ed., Cambridge University Press.
- SMITH, D.E., R. KOLENKIEWICZ, P.J. DUNN AND M.H. TORRENCE (1979). The measurement of fault motion by satellite laser ranging. *Proc. 6th International Symposium on Recent Crustal Movements*, Eds. C.A. Whitten, R. Green, B.K. Meade. Commission on Recent Crustal Movements, IAG, Stanford University, Palo Alto, U.S.A., July, 1977. *Tectonophysics* 52 (1-4), pp. 59-67.
- SNAY, R.A. (1976). Reducing the profile of sparse symmetric matrices. *Bull. Géod.* 50 (4), pp. 341-352.
- SODANO, E.M. (1965). General non-iterative solution of the inverse and direct geodetic problems. *Bull. Géod.* 75, pp. 69-89.
- STANLEY, H.R. (1979). The GEOS 3 project. *J. Geophys. Res.* 84 (B8), pp. 3779-3783.
- STOCH, L. (1963). Selecting stars for azimuth determination. *Bull. Géod.* 69, pp. 293-298.
- THOMAS, P.D. (1952). Conformal projections in geodesy and cartography. U.S. Coast and Geodetic Survey Special Publication 251, Washington, D.C., U.S.A.
- THOMAS, P.D. (1972). Long lines on the ellipsoid. Special publication of the National Ocean Survey of the NOAA, U.S. Department of Commerce, Washington, D.C., U.S.A.
- THOMPSON, M.M. (ED.) (1966). *Manual of Photogrammetry*. 3rd ed., Vol. I, American Society of Photogrammetry.
- THOMSON, D.B. (1976). Combination of geodetic networks. Department of Surveying Engineering Technical Report 30, University of New Brunswick, Fredericton, Canada.
- THOMSON, D.B. AND E.J. KRAKISKY (1976). Concepts of the combination of geodetic networks. *Proc. International Geodetic Symposium on Satellite Doppler Positioning*, DMA and NOS of the NOAA, New Mexico State University, Las Cruces, U.S.A., October. Physical Science Laboratory of the New Mexico State University, pp. 727-746.
- THOMSON, D.B. AND D.E. WELLS (1977). Hydrographic surveying I. Department of Surveying Engineering Lecture Note 45, University of New Brunswick, Fredericton, Canada.
- THOMSON, D.B., M.M. NASSAR AND C.L. MERRY (1974). Distortions of Canadian geodetic networks due to the neglect of deflections of the vertical and geoidal heights. *Canad. Surv.* 28 (5), pp. 598-605.

- THORSON, C.W. (1965). Second-order astronomical position determination manual. U.S. Coast and Geodetic Survey Publication 64-1, Washington, D.C., U.S.A.
- TOBEY, W.M. (1928). Geodesy. Geodetic Survey of Canada Publication 11, Department of Energy, Mines and Resources, Ottawa, Canada
- U.S. DEPARTMENT OF COMMERCE (1973). The North American datum. Publication of the National Ocean Survey of the NOAA, Rockville, U.S.A.
- U.S. FEDERAL GEODETIC CONTROL COMMITTEE (1974). Classification, standards of accuracy, and general specifications of geodetic control surveys. Report of the National Ocean Survey of the NOAA, Rockville, U.S.A.
- VAN DEN HOUT, C.M.A. (1966). The ANBLOCK method of planimetric block adjustment: mathematical foundation and organization of its practical application. *Photogrammetria* 21, pp. 171–178.
- VANIČEK, P. (1978). To the problem of noise reduction in sea level records used in vertical crustal movement detection. *Phys. of the Earth and Planetary Interiors* 17 (3), pp. 265–280.
- VANIČEK, P. AND G. CARERRA (1985). Reference ellipsoid misalignment, deflection components and geodetic azimuth. *Canad. Surv.*, 39(2), pp. 123–130.
- VANIČEK, P. AND E.W. GRAFARENDS (1980). On the weight estimation in levelling. NOAA Technical Report NOS 86 NGS 17, U.S. Department of Commerce, Rockville, U.S.A.
- VANIČEK, P. AND A.C. HAMILTON (1972). Further analysis of vertical crustal movement observations in the Lac St. Jean area, Québec. *Canad. J. Earth Sci.* 9 (9), pp. 1139–1147.
- VANIČEK, P. AND C.L. MERRY (1973). Determination of the geoid from deflections of the vertical using a least-squares surface fitting technique. *Bull. Géod.* 109, pp. 261–279.
- VANIČEK, P. AND D.E. WELLS (1974). Positioning of horizontal geodetic datums. *Canad. Surv.* 28 (5), pp. 531–538.
- VANIČEK, P., R.O. CASTLE AND E.I. BALAZS (1980). Geodetic leveling and its applications. *Rev. Geophys. and Space Physics* 18 (2), pp. 505–524.
- VANIČEK, P., K. THAPA AND D. SCHNEIDER (1981). The use of strain to identify incompatible observations and constraints in horizontal geodetic networks. *Manuscripta Geodaetica* VI (3), pp. 257–281.
- VANIČEK, P., R.B. LANGLEY, D.E. WELLS AND D. DELIKARAOGLOU (1984). Geometrical aspects of differential GPS positioning. *Bull. Géod.* 58, pp. 37–52.
- VEIS, G. (1960). Geodetic uses of artificial satellites. In: Smithsonian Contributions to Astrophysics 3 (9), pp. 95–161, Smithsonian Institution Astrophysical Observatory, Cambridge, U.S.A.
- VIGNAL, J. AND T.J. KUKKAMÄKI (1954). Comptes rendus des travaux de la section des nivelllements de précision. *Bull. Géod.*, supplement to Vol. 34, pp. 401–426.
- VINCENTY, T. (1973). Three-dimensional adjustment of geodetic networks. DMAAC Geodetic Survey Squadron, F.E. Warren AFB, Wyoming, U.S.A.
- VINCENTY, T. AND B.R. BOWRING (1978). Application of three-dimensional geodesy to adjustments of horizontal networks. NOAA Technical Memorandum NOS NGS-13, National Geodetic Survey, Rockville, U.S.A.
- VON ARX, W.S. (1967). Relationship between marine geodesy and oceanographic measurements. *Proc. 1st Marine Geodesy Symposium*, Battelle Memorial Institute, Columbus, U.S.A., September, 1966. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 37–42.
- WASSEF, A.M. AND F.Z.A. MESSIH (1960). On the statistical distribution of levelling errors. *Bull. Géod.* 55, pp. 201–210.
- WELLS, D.E. (1974). Doppler satellite control. Department of Surveying Engineering Technical Report 29, University of New Brunswick, Fredericton, Canada.
- WELLS, D.E. AND S. GRANT (1977). Reliable navigation through system integration. *Proc. 16th Annual Canadian Hydrographic Conference*, CHS and CHA, Burlington, Canada, May. Special ed. of *Lighthouse*, Journal of the CHA.
- WELLS, D.E., AND P. VANIČEK (1975). Alignment of geodetic and satellite coordinate systems to the average terrestrial system. *Bull. Géod.* 117, pp. 241–257.
- WELLS, D.E., D. DELIKARAOGLOU AND P. VANIČEK (1982). Marine navigation with NAVSTAR/Global Positioning System (GPS) today and in the future. *Canad. Surv.* 36(1), pp. 9–28.
- WELLS, D.E., P. VANIČEK AND D. DELIKARAOGLOU (1981). Application of NAVSTAR/GPS to geodesy in Canada: Pilot study. Department of Surveying Engineering Technical Report 76, University of New Brunswick, Fredericton, Canada.

- WELLS, D.E., E.J. KRAKIWSKY, D.B. THOMSON AND J. KOUBA (1976). Review of Doppler satellite positioning in Canada. *Bull. Géod.* 50 (4), pp. 307-321.
- WESTERFIELD, E.E. AND G. WORSLEY (1966). Translocation by navigation satellite. *APL Tech. Dig.* 5 (5), pp. 2-10.
- WISE, P.J. (1979). Laser terrain profiler. Division of National Mapping Technical Report 26, Department of National Development, Canberra, Australia.
- WOLF, H. (1978). The Helmert block method—its origin and development. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 319-326.
- WOLF, P. (1974). *Elements of Photogrammetry*. McGraw-Hill.
- YEREMEYEV, V.F. AND M.I. YURKINA (1969). On orientation of the reference geodetic ellipsoid. *Bull. Géod.* 91, pp. 13-15.
- YIONOULIS, S.M. (1970). Algorithm to compute tropospheric refraction effects on range measurements. *J. Geophys. Res.* 75, pp. 7636-7637.
- ZAKATOV, P.S. (1953). *A Course in Higher Geodesy*. Translated from Russian by the Israel Program for Scientific Translations for the National Science Foundation and the Department of Commerce, Washington, D.C., U.S.A., 1962.

PART V

EARTH'S GRAVITY FIELD

## CHAPTER 20

# GLOBAL TREATMENT OF THE GRAVITY FIELD

The aim of this chapter is to develop the mathematical apparatus needed for handling the earth's gravity field in a global sense. The four sections correspond to the four standard components of this apparatus. In the first section, the basic boundary value problem for gravity potential is developed simultaneously with an alternative approach using integral equations. The second section discusses the selection of appropriate coordinate systems and shows how the classical method of separation of variables can be used to express the gravitational potential in the form of an infinite series. In the third section, the reference system covered in Chapter 6—the normal gravity field—is revisited. This time, however, details are given on how this system is established. The fourth section is devoted to the definition of a very important quantity—the disturbing potential—the introduction of which allows us to linearize, in a natural way, many of the mathematical models used later in this part.

### 20.1. Fundamental equations for gravity potential

As was shown in §6.3, the (vector) gravity field  $\bar{g}$  can be completely and uniquely represented by a scalar field, the gravity potential  $W$ . This being the case, once the potential  $W$  is known in the region of interest, all the other parameters characterizing the behaviour of the gravity field can be derived from it. Hence we shall concentrate here on the ways of obtaining  $W$ , to the exclusion of other parameters; the technique of converting  $W$ —or, more accurately, of converting only its irregular part, the disturbing potential—into other parameters will be treated in §21.1.

To derive the fundamental partial differential equation that describes the behaviour of the gravity potential  $W$ , let us first have a look at the local behaviour of the gravitational vector  $\bar{g}_g$  (cf. §6.1). Its behaviour at a point is fully described by its curl and divergence (cf. §3.2) given in the neighbourhood of that point.

Since the gravitational potential field is irrotational and, in the first approximation, conservative (cf. §6.3), it means that we can immediately write

$$\mathbf{curl} \nabla W_g(\bar{r}_A) = \mathbf{curl} \bar{g}_g(\bar{r}_A) = \nabla \times \bar{g}_g(\bar{r}_A) = \bar{0} \quad (20.1)$$

for any point  $A$ , where  $\bar{r}_A$  is a position vector in an arbitrary system of coordinates.

On the other hand, the divergence of  $\bar{g}_g(\vec{r}_A)$  can be written as the limiting case of the Gauss formula (3.54): i.e.,

$$\operatorname{div} \bar{g}_g(\vec{r}_A) = \nabla \cdot \bar{g}_g(\vec{r}_A) = \lim_{V_A \rightarrow 0} \frac{\iint_{\mathcal{S}} \bar{g}_g(\vec{r}) \cdot \vec{n} dS}{V_A}, \quad (20.2)$$

where  $V_A$  is the volume enclosed by the surface  $\mathcal{S}$ , and  $\vec{n}$  is a unit outward vector normal to  $\mathcal{S}$  (see FIG. 1). If the integral of the *gravitational flux*  $\bar{g}_g \cdot \vec{n} dS$  is positive, point  $A$  is called a *source*, if the integral is negative, then  $A$  is called a *sink*. In the case of the integral being zero,  $\bar{g}_g$  is not divergent at  $A$ . At the outset, this is the important question: Is the divergence of the gravitational acceleration positive, negative, or zero?

To work out the answer, let us first denote the mass  $M$  of volume  $V_A$  embraced by the surface  $\mathcal{S}$  as

$$M = V_A \sigma(\vec{r}_A), \quad (20.3)$$

with the understanding that for a sufficiently small volume  $V_A$  the mass density  $\sigma$  is uniform (constant) in  $V_A$ . If, for simplicity, the origin of the coordinate system is put at the centre of  $V_A$ , we discover that  $V_A$  generates an acceleration  $\bar{g}_g$  given by

$$\bar{g}_g(\vec{r}_A) = \frac{GM}{|\vec{r}_A|^3} \vec{r}_A \quad (20.4)$$

(cf. (6.2) and (6.10)). It can further be shown that when the limit is considered in (2), the shape of the surface  $\mathcal{S}$  is immaterial as long as it fully encloses the volume  $V_A$ . Thus, again for simplicity, a spherical surface can be chosen. In this case, it is clear that the  $\bar{g}_g$  generated by the spherical volume  $V_A$  is normal to  $\mathcal{S}$  everywhere, and  $\bar{g}_g \cdot \vec{n}$  is equal to  $g_g$ . The gravitational flux can then be written as

$$\bar{g}_g \cdot \vec{n} dS = g_g dS = - \frac{GM}{|\vec{r}|^2} dS = - \frac{G\sigma(\vec{r}_A)V_A}{|\vec{r}|^2} dS, \quad (20.5)$$

where  $r$  is the radius of the sphere  $\mathcal{S}$ .

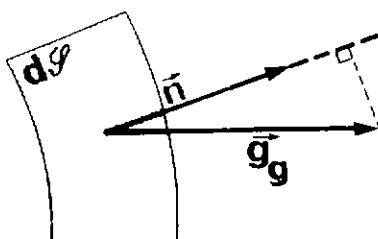


FIG. 20.1 Gravitational flux.

With the realization that the surface of a sphere of radius  $r$  is equal to  $4\pi r^2$ , it is not difficult to see that the integral of the gravitational flux gives

$$\iint_S \bar{g}_g \cdot \bar{n} dS = -4\pi G \sigma(\bar{r}_A) V_A, \quad (20.6)$$

and that upon dividing by  $V_A$  and taking the limit we obtain

$$\operatorname{div} \bar{g}_g(\bar{r}_A) = -4\pi G \sigma(\bar{r}_A). \quad (20.7)$$

This equation, which can also be obtained directly from the definition of divergence by taking the limiting case of  $r \rightarrow 0$ , is valid, of course, in any coordinate system. Its interpretation is as follows: since  $\sigma(\bar{r}_A)$  is a non-negative quantity,  $A$  can be either a sink of gravitational acceleration  $\bar{g}_g(\bar{r}_A)$  (if  $\sigma(\bar{r}_A) > 0$ ) or else the divergence of  $\bar{g}_g(\bar{r}_A)$  is zero in  $A$  (if  $\sigma(\bar{r}_A) = 0$ ).

Having disposed of the local behaviour of gravitational acceleration, let us now turn to centrifugal acceleration  $\bar{g}_c(\bar{r}_A)$  (cf. §6.1). Once more, the centrifugal field is irrotational, and

$$\operatorname{curl} \bar{g}_c(\bar{r}_A) = \bar{0}. \quad (20.8)$$

The divergence of  $\bar{g}_c(\bar{r}_A)$  is easily evaluated by selecting the coordinate system  $(x, y, z)$  so that the  $z$ -axis coincides with the spin axis of the earth. Recalling (6.7) and realizing that  $p = \sqrt{x^2 + y^2} \doteq a \cos \phi$ , we get

$$\bar{g}_c(\bar{r}_A) = \omega^2 (x_A \bar{i} + y_A \bar{j}), \quad (20.9)$$

and the divergence is readily obtained as

$$\operatorname{div} \bar{g}_c(\bar{r}_A) = 2\omega^2, \quad (20.10)$$

following the rules of vector analysis (cf. §3.2). The value of  $\operatorname{div} \bar{g}_c$ , being a scalar field, must not change with transformation into any other coordinate system, since the behaviour of  $\bar{g}_c$  does not depend on the coordinate system. Hence we discover that the divergence of centrifugal acceleration remains constant ( $2\omega^2$ ) throughout the space.

Realizing that gravity acceleration  $\bar{g}(\bar{r}_A)$  is the sum of the two accelerations treated above, we obtain, as a consequence of the linearity of the  $\nabla$  operator,

$$\nabla \cdot \bar{g}(\bar{r}_A) = \operatorname{div} \bar{g}(\bar{r}_A) = -4\pi G \sigma(\bar{r}_A) + 2\omega^2. \quad (20.11)$$

Further, since  $\bar{g} = \nabla W$ , we finally get

$$\boxed{\nabla^2 W(\bar{r}_A) = -4\pi G \sigma(\bar{r}_A) + 2\omega^2.} \quad (20.12)$$

This is the fundamental, partial differential equation (of second order) for gravity potential that we set out to derive. It is recognized as being of the Poisson type (cf. §3.2).

Although (12) is valid in this form throughout the space, it is sometimes expedient to spell out its two special forms:

(a) One applies in empty space ( $\sigma = 0$ ) and reads

$$\nabla^2 W(\bar{r}) = 2\omega^2. \quad (20.13)$$

Disregarding the density of the atmosphere (which is negligibly small compared with the density of the earth, cf. §9.1), we can say that (13) is valid outside the earth.

(b) The second special case of (12) can be formulated for the surface of the earth. A point  $A$  located on the surface of the earth may be considered as having a density equal to a fraction of the density at a point which is entirely within the earth. In other words, any differential neighbourhood of point  $A$  on the earth's surface is partially empty and partially dense. Hence,

$$\nabla^2 W(\bar{r}_A) = -k\pi G\sigma(\bar{r}_A) + 2\omega^2, \quad (20.14)$$

where  $\sigma(\bar{r}_A)$  is understood to be the density of the material at the earth's surface, and  $k \in (0, 4)$ —see FIG. 2.

These two equations illustrate the fact that the second derivatives of gravity potential (contained in the left-hand sides) are discontinuous on a boundary between any two media of different densities, the earth's surface being a notable example. This fact will be extensively employed in Chapter 21. It may also be noted that the fundamental differential equation for gravitational potential  $W_g$  outside the earth (strictly speaking, in empty space), known as the *potential equation*, is of the Laplace type:

$$\boxed{\nabla^2 W_g = 0.} \quad (20.15)$$

The gravitational potential in empty space is a harmonic function (cf. §3.2).

Clearly, the problem of determining the earth's gravity field can be broken down to the determination of the gravitational field and the determination of the centrifugal field. While the former task is indeed formidable, the determination of the centrifugal field is not:  $\bar{g}_c$  is a simple function of position since the earth-spin angular velocity is known very accurately (cf. §5.2). Consequently, in the remainder of this section, as well as in most of the rest of this part, we shall concentrate on the gravitational field.

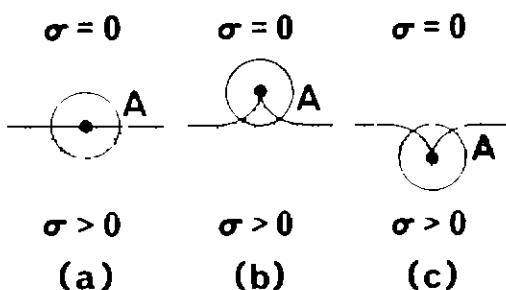


FIG. 20.2. Point on the earth's surface. (a)  $k = 2$ ; (b)  $0 < k < 2$ ; (c)  $2 < k < 4$  (according to KOUBA [1979]).

Geodetic interest lies mainly in the space immediately outside the earth (to an altitude of a few thousand kilometres), the earth's surface, and the uppermost layer of the earth's crust. From the character of the differential equations valid within and outside the earth, one can see that it is expedient to split the region of interest into two: outside and inside the earth. Accordingly,  $W_g$  outside the earth is often spoken of as the earth's *external gravitational potential* and inside the earth as the *internal gravitational potential*. At first glance, the determination of the earth's gravitational potential  $W_g$ , at least outside the earth, should not be very difficult. External  $W_g$  satisfies the Laplace equation outside the earth, and it should thus be enough to obtain the values of the gravitational potential  $W_g$  (in one form or another) on the surface of the earth, i.e., on the boundary of the region of interest, and solve for  $W_g$  outside the earth using a technique for solving the Laplace boundary value problem (cf. §3.2). The external gravitational potential is the one we will concentrate on in this book.

Unfortunately, the earth's surface is not sufficiently smooth to ensure the uniqueness of such an external solution (see §3.2). Possible ways of overcoming this obstacle will be discussed in Chapter 22, together with ways of setting up the boundary values. Let it suffice, for the present, to acknowledge that it can be done, and turn our attention to the available techniques for solving the boundary value problem. From among the many existing techniques (cf. §3.2), we shall concentrate on the two that are most readily applicable here: transformation to integral equations, and the method of separation of variables (the Fourier technique).

To formulate the integral equation corresponding to our boundary value problem, let us first invoke Green's second identity (3.57) and take  $Q(\vec{r}) = W_g(\vec{r})$  and  $P(\vec{r}_A) = 1/|\vec{r} - \vec{r}_A| = 1/\rho$  (see FIG. 3). The reader can easily verify that the second function is actually a kernel, i.e.,  $P(\vec{r}_A, \vec{r}) = 1/\rho$ , and is harmonic everywhere except for  $\rho = 0$ . It is called the *fundamental harmonic function*. Consequently, we obtain

$$\begin{aligned} & \oint_S \left[ \frac{1}{\rho} \frac{\partial W_g(\vec{r})}{\partial n} - W_g(\vec{r}) \frac{\partial}{\partial n} \frac{1}{\rho} \right] dS - \iiint_B \frac{1}{\rho} \nabla^2 W_g(\vec{r}) d\mathcal{B} \\ &= - \iiint_B W_g(\vec{r}) \nabla^2 \left( \frac{1}{\rho} \right) d\mathcal{B}, \end{aligned} \quad (20.16)$$

which is known as Green's third identity. It can be simplified as follows: first, note that the gravitational potential is taken only within the earth and, as such, is a

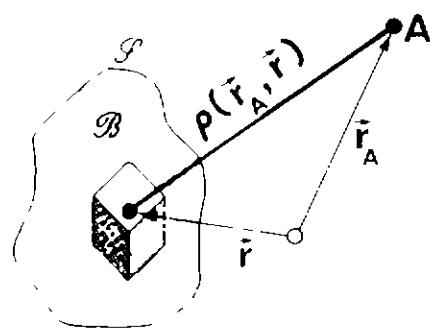


FIG. 20.3. Fundamental harmonic function.

function of the dummy variable  $\bar{r}$  only. On the other hand, since the fundamental harmonic function is a kernel (cf. §3.2), it is a function of both  $\bar{r}$  and  $\bar{r}_A$ . Depending on the position of the point  $A$ , the right-hand side of (16) is equal to  $4\pi W_g(\bar{r}_A)$  for  $A$  inside  $\mathcal{S}$ ,  $2\pi W_g(\bar{r}_A)$  for  $A$  on smooth  $\mathcal{S}$ , and 0 for  $A$  outside  $\mathcal{S}$  [MACMILLAN, 1930]. Hence, Green's third identity becomes

$$\begin{aligned} \oint_{\mathcal{S}} \left( \frac{1}{\rho} \frac{\partial W_g}{\partial n} - W_g \frac{\partial \rho^{-1}}{\partial n} \right) d\mathcal{S} - \iiint_{\mathcal{B}} \frac{1}{\rho} \nabla^2 W_g d\mathcal{B} = \\ = \begin{cases} 4\pi W_g(\bar{r}_A), & A \text{ inside } \mathcal{S}, \\ 2\pi W_g(\bar{r}_A), & A \text{ on smooth } \mathcal{S}, \\ 0, & A \text{ outside } \mathcal{S}. \end{cases} \end{aligned} \quad (20.17)$$

With the realization now that  $\nabla^2 W_g(\bar{r}) = -4\pi G\sigma(\bar{r})$  inside the earth, the volume integral becomes  $4\pi G \iiint_{\mathcal{B}} [\sigma(\bar{r})/\rho] d\mathcal{B}$ , which is simply  $4\pi W_g(\bar{r}_A)$  (cf. (6.25)). Thus we finally get

$$\oint_{\mathcal{S}} \left( \frac{1}{\rho} \frac{\partial W_g}{\partial n} - W_g \frac{\partial \rho^{-1}}{\partial n} \right) d\mathcal{S} = \begin{cases} 0, & A \text{ inside } \mathcal{S}, \\ -2\pi W_g(\bar{r}_A), & A \text{ on smooth } \mathcal{S}, \\ -4\pi W_g(\bar{r}_A), & A \text{ outside } \mathcal{S}. \end{cases} \quad (20.18)$$

This is the solution of the Poisson boundary value problem for  $W_g$  in integral form. In particular, outside the earth ( $A$  outside  $\mathcal{S}$ ) the above integral equation is equivalent to the Laplace boundary value problem. Again, however, our earlier remark on the insufficient smoothness of boundary  $\mathcal{S}$  is valid here; the application of this approach will be treated in Chapter 22.

The next technique for solving the boundary value problem is the separation of variables (see §3.2). To obtain the solution in its simplest form, it is expedient to select a more convenient coordinate system than the Cartesian. Two such systems are commonly being used—the spherical and the ellipsoidal. Section 3.2 showed that the form of the Laplacean  $\nabla^2$  is different in different coordinate systems; in spherical coordinates  $(r, \theta, \lambda)$ , eqn. (15) acquires the following form:

$$2r \frac{\partial W_g}{\partial r} + r^2 \frac{\partial^2 W_g}{\partial r^2} + \cot \theta \frac{\partial W_g}{\partial \theta} + \frac{\partial^2 W_g}{\partial \theta^2} + \sin^{-2} \theta \frac{\partial^2 W_g}{\partial \lambda^2} = 0. \quad (20.19)$$

The approach using an ellipsoidal coordinate system is more complex. Conceivably, geodetic coordinates  $(\phi, \lambda, h)$  could be used here. However, as was indicated in §3.3, they represent a two-parametric system and consequently render the expressions too cumbersome. For this reason, a *one-parametric ellipsoidal system of coordinates* (EL) is preferred in this context. Relatively simple expressions are obtained for the ellipsoidal system that uses the focal length  $E$ ,

$$E = \sqrt{a^2 - b^2} = ae, \quad (20.20)$$

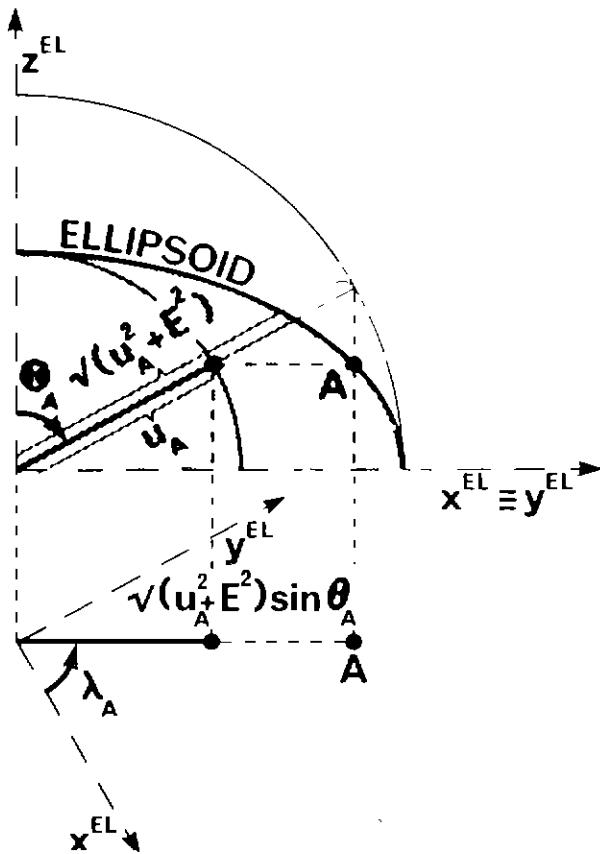


FIG. 20.4. One-parametric ellipsoidal coordinate system.

as the parameter. The three coordinates  $u, \Theta, \lambda$  are shown in FIG. 4. Note that when the parameter  $E$  is held fixed in this system, points with different  $u$ -coordinates lie on different ellipsoids.

To familiarize ourselves with this one-parametric ellipsoidal system, let us derive the transformation equations to and from the representative Cartesian system (see §3.3). They are immediately obtained from FIG. 4:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}^{\text{EL}} = \begin{bmatrix} \sqrt{u^2 + E^2} \sin \Theta \cos \lambda \\ \sqrt{u^2 + E^2} \sin \Theta \sin \lambda \\ u \cos \Theta \end{bmatrix}. \quad (20.21)$$

Inverse transformations can be derived from the above equations yielding (leaving out the superscript EL)

$$\begin{bmatrix} u \\ \Theta \\ \lambda \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{r^2 - E^2}{2}} \sqrt{1 + \sqrt{1 + \frac{4z^2 E^2}{(r^2 - E^2)^2}}} \\ \arccos(z/u) \\ \arctan(y/x) \end{bmatrix}. \quad (20.22)$$

where  $r = \sqrt{x^2 + y^2 + z^2}$ , and  $u$  is, for simplicity, used in the second equation. The derivation is left to the reader.

It is also left to the reader to show that transformation equations to and from the spherical system of the same family are

$$\begin{bmatrix} r \\ \theta \\ \lambda \end{bmatrix} = \begin{bmatrix} \sqrt{u^2 + E^2 \sin^2 \Theta} \\ \arccos((u/r) \cos \Theta) \\ \lambda \end{bmatrix}, \quad (20.23)$$

and

$$\begin{bmatrix} u \\ \Theta \\ \lambda \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{r^2 - E^2}{2}} \sqrt{1 + \sqrt{1 + \frac{4r^2 E^2 \cos^2 \theta}{(r^2 - E^2)^2}}} \\ \arccos((r/u) \cos \theta) \\ \lambda \end{bmatrix} \quad (20.24)$$

where, again for simplicity,  $u$  and  $r$  are left in the second equations. The relations between geodetic and ellipsoidal coordinates are more involved, and the transformations are best done through the Cartesian coordinate system. Let us just state here that if the point  $P$  lies on the reference ellipsoid  $(a, b)$  of the geodetic system, then  $u = b$ ,  $\sqrt{u^2 + E^2} = a$ , and it can be shown that [BOMFORD, 1971]

$$\tan \Theta = \frac{a}{b} \cot \phi, \quad (20.25)$$

and, again, the longitudes  $\lambda$  are identical.

The Laplace equation for  $W_g$  in ellipsoidal coordinates has been derived, for instance, by HEISKANEN AND MORITZ [1967]. Only their result is shown here:

$$2u \frac{\partial W_g}{\partial u} + (u^2 + E^2) \frac{\partial^2 W_g}{\partial u^2} + \cot \Theta \frac{\partial W_g}{\partial \Theta} + \frac{\partial^2 W_g}{\partial \Theta^2} + \frac{u^2 + E^2 \cos^2 \Theta}{(u^2 + E^2) \sin^2 \Theta} \frac{\partial^2 W_g}{\partial \lambda^2} = 0.$$

$$(20.26)$$

Equations (19) and (26) are the two fundamental differential equations for the earth's external gravitational potential in use here; they are equivalent to the third equation (18). Note that if the EL system is defined so that  $E = 0$ , it becomes identical with the spherical system:  $u$  becomes  $r$ ,  $\Theta$  becomes  $\theta$ , and, indeed, (26) transforms into (19). Thus, to an accuracy of  $e^2$ , the ellipsoidal coordinates can be replaced by the spherical in any expression.

We are now ready to apply the method of separation of variables to (19) and (26). Since this step is very important, the entire next section has been devoted to it.

## 20.2. Eigenfunction development of gravitational potential

When applying the method of separation of variables (see §3.2), (19) should be dealt with first and the gravitational potential  $W_g(r, \theta, \lambda)$  sought in terms of a product of the following three functions:

$$W_g(r, \theta, \lambda) = R(r) \cdot T(\theta) \cdot L(\lambda). \quad (20.27)$$

It will be useful to denote the product of the last two functions by  $J(\theta, \lambda)$ .

If the spherical coordinate system  $(r, \theta, \lambda)$  is selected so that the origin is close to the centre of mass of the earth, the surface of the earth will be very close to the sphere  $r = R$ . Then the first function,  $R$ , will describe the behaviour of  $W_g$  in the direction approximately normal to the earth's surface (globally speaking), while the function  $J$  will depict the variations of the potential on the earth's surface. More precisely, if we take a sphere  $S$  of radius  $r = a$ , called the *Brillouin sphere* (see FIG. 5), that encompasses all the masses of the earth, then on this sphere  $R$  will be constant, and all the variations of  $W_g$  on  $S$  will be characterized by  $J$ . Moreover, for the purpose of this section, this sphere can be taken as the lower boundary for the external boundary value problem, and the upper boundary for the internal problem. Clearly, disregarding the atmosphere, (19) is valid outside this sphere.

The first application of the Fourier method to (19) yields

$$r^2 R'' + 2rR' - c_1 R = 0, \quad (20.28)$$

and

$$\cot \theta \frac{\partial J}{\partial \theta} + \frac{\partial^2 J}{\partial \theta^2} + \sin^{-2} \theta \frac{\partial^2 J}{\partial \lambda^2} + c_1 J = 0. \quad (20.29)$$

Equation (28) is an ordinary differential equation of second order; the primes denote derivatives with respect to  $r$ . Equation (29) is still a partial differential equation in  $\theta$  and  $\lambda$ . The second application of the Fourier method (to (29)) yields:

$$\sin^2 \theta T'' + \sin \theta \cos \theta T' + (c_1 \sin^2 \theta - c_2) T = 0, \quad (20.30)$$

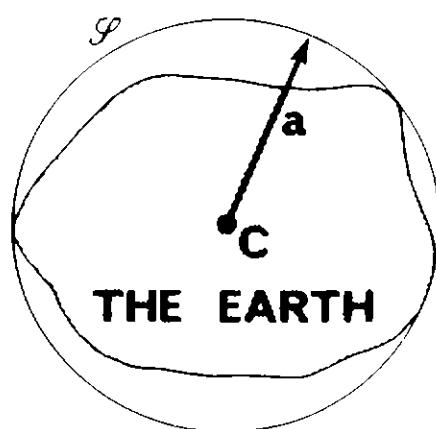


FIG. 20.5. Brillouin's sphere.

and

$$L'' + c_2 L = 0. \quad (20.31)$$

In these two equations, the variables  $\theta$  and  $\lambda$  are separated. The system of the three ordinary differential equations, i.e., (28), (30), and (31), is equivalent to (19)—see §3.2. Their interrelation is ensured through the two common constants  $c_1, c_2$ . The three ordinary differential equations give the general solution for the functions  $R, T$ , and  $L$ .

Let us first tackle the last two equations. Equation (31) is readily recognized as an equation of simple harmonic motion whose eigenvalues  $m = \sqrt{c_2}$  and eigenfunctions were discussed in §3.2. Any linear combination  $L$  of these eigenfunctions satisfies (31).

Substituting  $t$  for  $\cos \theta$  and  $m^2$  for  $c_2$  in (30), we obtain:

$$(1-t^2)T'' - 2tT' + \left( c_1 - \frac{m^2}{1-t^2} \right) T = 0, \quad (20.32)$$

where the derivatives are now taken with respect to  $t$ . This is a second-order Legendre equation for  $T$  (cf. §3.2). It makes sense to attempt its solution only for such values of  $m$  for which we know the solution of (31) exists, i.e., only for the eigenvalues of (31). It has been shown by, e.g., HOBSON [1931] that (32) gives a solution only for the following eigenvalues  $c_1$ :

$$c_1 = n(n+1), \quad n = m, m+1, m+2, \dots, \quad (20.33)$$

when  $\theta \in (0, \pi)$ . These are the admissible eigenvalues of the Legendre equation. Corresponding eigenfunctions are the *Legendre associated functions* (of  $n$ th degree and  $m$ th order) given by

$$P_{nm}(t) = (1-t^2)^{m/2} \frac{d^m}{dt^m} P_n(t), \quad (20.34)$$

where

$$P_n(t) = \frac{1}{n!2^n} \frac{d^n}{dt^n} (t^2 - 1)^n \quad (20.35)$$

are the Legendre functions already seen in §3.2 and §8.1. Their properties are discussed in detail, for instance, in ABRAMOWITZ AND STEGUN [1964]; it is to be noted here that again the eigenfunctions are orthogonal for  $t \in (-1, 1)$ , i.e.,  $\theta \in (0, \pi)$ , with weight  $w(t) = 1$ .

Getting back to (29), we can see that it is satisfied by any function  $J$  that is a product of a linear combination of trigonometric functions, with a linear combination of eigenfunctions (34) for the admissible values of  $m$  and  $n$ . Such a product can generally be written as

$$\begin{aligned} J(\theta, \lambda) &= \sum_{n=0}^{\infty} \sum_{m=0}^n (A_{nm} \cos m\lambda + B_{nm} \sin m\lambda) P_{nm}(\cos \theta) \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^n (A_{nm} Y_{nm}^c + B_{nm} Y_{nm}^s), \end{aligned} \quad (20.36)$$

where  $A_{nm}$ ,  $B_{nm}$  are some arbitrary numbers. The functions  $Y_{nm}^c = \cos m\lambda P_{nm}(\cos \theta)$  and  $Y_{nm}^s = \sin m\lambda P_{nm}(\cos \theta)$  are called (surface) *spherical harmonic functions*. They can be regarded as eigenfunctions of the Laplace equation in spherical coordinates on the surface of a sphere. Integers  $m = 0, 1, 2, \dots$ ;  $n = m, m+1, m+2, \dots$ , or, identically,  $n = 0, 1, 2, \dots$ ;  $m = 0, 1, \dots, n$ , are then the eigenvalues of the Laplace equation.

Note once more that the spherical harmonics, being eigenfunctions, are indeed orthogonal on a spherical surface  $S$ , i.e., for any  $(\theta, \lambda) \in (0, \pi) \times (-\pi, \pi)$ . This means that the following equation,

$$\oint_S Y_{nm}^k(\theta, \lambda) Y_{ij}'(\theta, \lambda) d\nu = 0, \quad (20.37)$$

where  $d\nu$  is the solid angle element, is satisfied for  $i \neq n$  or  $j \neq m$  or  $k \neq l$ . The solid angle rather than the surface element is used to give a unitless result of the integration. Moreover, it can be shown that their norms (cf. (3.61) and §11.1) are equal to [HOBSON, 1931], for both superscripts c and s,

$$\|Y_{nm}\|^2 = \oint_S Y_{nm}^2 d\nu = \begin{cases} \frac{4\pi}{2n+1}, & m=0, \\ \frac{2\pi}{2n+1} \frac{(n+m)!}{(n-m)!}, & m \neq 0. \end{cases} \quad (20.38)$$

Functions  $\tilde{Y}_{nm} = Y_{nm}/\|Y_{nm}\|$  are then orthonormal, i.e.,

$$\|\tilde{Y}_{nm}\| = 1, \quad (20.39)$$

and are thus called (fully) *normalized spherical harmonic functions*. Spherical harmonic functions for  $m=0$  are called *zonal harmonics* and the other functions *tesseral harmonics*.

Returning to the main problem at hand, the exterior problem for  $W_g$  (outside the Brillouin sphere), we can now use spherical harmonics to solve it quite easily. Considering (27) and (36), the solution is clearly given as

$$W_g(r, \theta, \lambda) = R(r) \sum_{n=0}^{\infty} \sum_{m=0}^n (A_{nm} Y_{nm}^c(\theta, \lambda) + B_{nm} Y_{nm}^s(\theta, \lambda)). \quad (20.40)$$

Here, the coefficients  $A_{nm}$ ,  $B_{nm}$  have to be determined in such a way as to satisfy the boundary value  $W_g(r=a, \theta, \lambda)$  on the sphere. That means the following equation would have to be satisfied:

$$W_g(a, \theta, \lambda) = R(a) \sum_{n=0}^{\infty} \sum_{m=0}^n (A_{nm} Y_{nm}^c(\theta, \lambda) + B_{nm} Y_{nm}^s(\theta, \lambda)), \quad (20.41)$$

where, of course,  $R(a)$  is constant.

We are now ready to focus our attention on the third ordinary differential equation (28). Here again, it is meaningful to seek the solution only for the admissible values of  $c_1 = n(n+1)$ . It is a variety of an *Euler equation* and is readily solved using substitution of  $\exp(it)$  for  $r$ . We find that two families of functions,  $R_1$  and  $R_2$ ,

$$R_1(r) = \exp(nt) = r^n, \quad R_2(r) = \exp((-n-1)t) = r^{-(n+1)}, \quad (20.42)$$

satisfy the equation [HOCHSTADT, 1964] for any admissible  $n$ . The first functions increase beyond limits when  $r$  grows beyond limits. This is inconsistent with the expected physical behaviour of the gravitational potential (cf. §6.3). Hence, when a solution to the external problem is sought, we are obliged to take the second family of functions to make the solution compatible with physical requirements. For parallel reasons,  $R_1$  is used when the Laplace interior problem is contemplated.

It is expedient to make the radial function  $R = R_2$  unitless so that only the coefficients  $A_{nm}$ ,  $B_{nm}$  in (40) are the carriers of physical units. This is normally done by scaling the spherical coordinate system so as to make the radius of the boundary sphere a unity: instead of using  $r$ , we use  $r/a$ . In this coordinate system, the radial functions become

$$\left(\frac{r}{a}\right)^{-(n+1)} = \left(\frac{a}{r}\right)^{n+1}, \quad n=0, 1, \dots \quad (20.43)$$

Such a scaling, of course, affects even the values of the coefficients in (40) and thus preserves the correct overall magnitude of  $W_g$ .

Let us now rewrite (40) in its final form. For the external gravitational potential, we have

$$W_g(r, \theta, \lambda) = \sum_{n=0}^{\infty} \left(\frac{a}{r}\right)^{n+1} \sum_{m=0}^n (A_{nm} \cos m\lambda + B_{nm} \sin m\lambda) P_{nm}(\cos \theta).$$

(20.44)

This equation is equivalent to (6.25), the only difference being that instead of the unknown mass density  $\sigma$ , here we have infinitely many unknown coefficients  $A_{nm}$  and  $B_{nm}$  (in the units of potential). Equation (44) is known as the *development of gravitational potential into spherical harmonics*, or development into the eigenfunctions of the Laplace equation in spherical coordinates.

It is illustrative to realize that the shape of the equipotential surfaces of  $W_g$ , represented by (44), depends on the dominance of the appropriate terms in the series. An equatorial bulge is caused by the presence of the  $(2,0)$  term; the modelling of the bulge by the zonal harmonic  $P_{2,0}$  is shown in FIG. 6(a). The pear-shapedness is caused by the presence of the  $(3,0)$  term—see FIG. 6(b).

The tesseral harmonics can be regarded as being created by a meridian profile that is then modulated longitudinally. Denoting  $(A_{nm}^2 + B_{nm}^2)$  by  $R_{nm}^2$  and  $B_{nm}/A_{nm}$  by  $\tan \kappa_{nm}$ , we can write

$$(A_{nm} \cos m\lambda + B_{nm} \sin m\lambda) P_{nm}(\cos \theta) = R_{nm} P_{nm}(\cos \theta) \cos(m\lambda - \kappa_{nm}). \quad (20.45)$$

Clearly, the basic meridian component  $R_{nm} P_{nm}(\cos \theta)$  is modulated by  $\cos(m\lambda - \kappa_{nm})$ . FIG. 7 illustrates a case of  $m=2$ .

Before getting into the realm of determination of the unknown coefficients, let us reflect for a moment on the role of the radial terms  $(a/r)^{n+1}$ . Outside the boundary

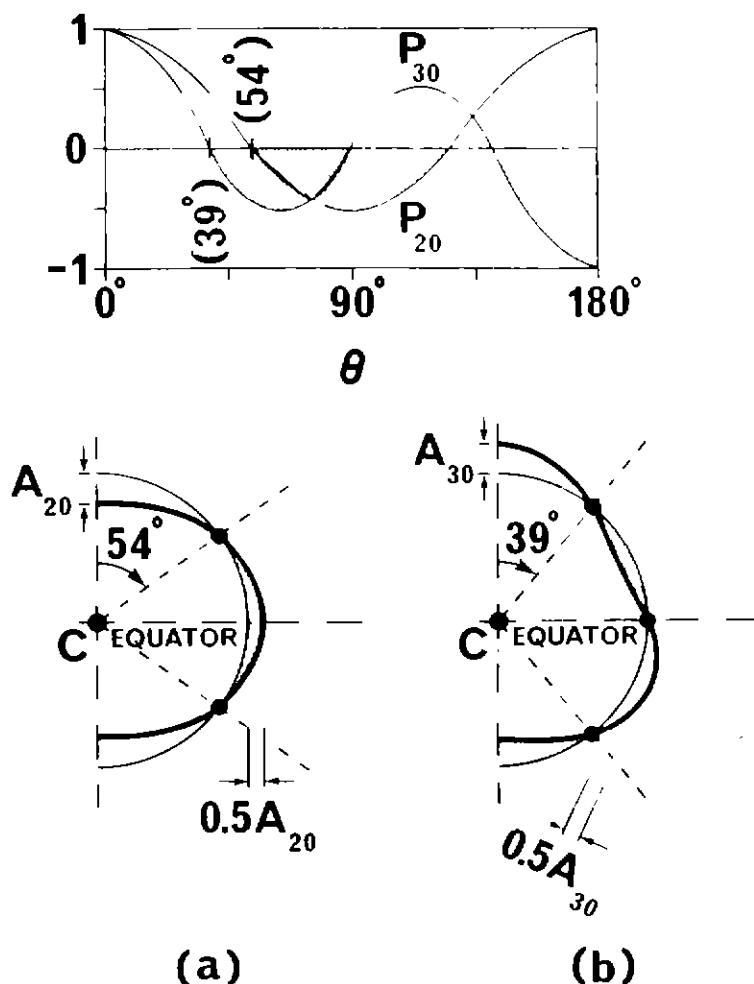


FIG. 20.6. Modelling of the equatorial bulge and pear-shapedness.

sphere, their magnitudes diminish with increasing order  $n$ , since  $a/r < 1$ . This means that with increasing altitude, the higher frequency ( $n$  and  $m$ ) wavelength features of the potential field tend to be smoothed out. This is a very important fact, and its repercussions will be discussed in Chapter 23.

The question that may linger in the reader's mind is what has been gained by deriving (44) instead of (6.25), since we have succeeded only in replacing one unknown function of location (namely  $\sigma$ ) by a whole series of unknowns (the coefficients). In fact, we have progressed at least one step farther, because we are now able to evaluate the coefficients from data on and above the earth instead of being dependent on the knowledge of mass distribution inside the earth. If the value of the gravitational potential on the boundary sphere  $S$  ( $r = a$ ) is known, i.e., if

$$V(\theta, \lambda) = W_g(a, \theta, \lambda) \quad (20.46)$$

is known, then the *potential coefficients*  $A_{nm}$ ,  $B_{nm}$  could be obtained through the standard procedure of developing  $V$  into spherical harmonics. The equations for potential coefficients, given here for simplicity when normalized spherical harmonic

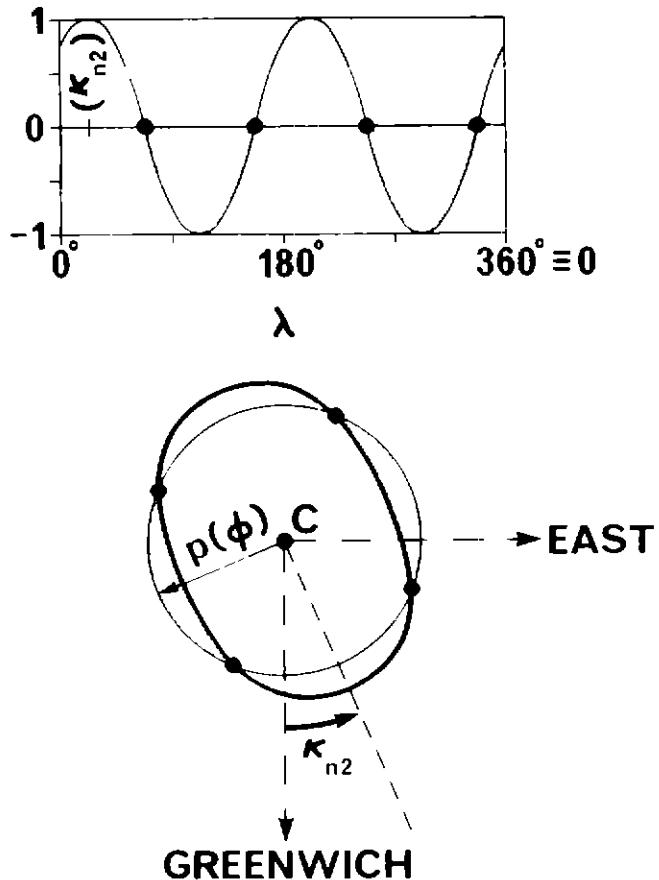


FIG. 20.7. Modelling of longitudinal variation through modulation by  $\cos(2\lambda - \kappa_{n2})$ .

functions are used [HOBSON, 1931], are:

$$\begin{aligned}\tilde{A}_{nm} &= \iint_S V(\theta, \lambda) \tilde{Y}_{nm}^c(\theta, \lambda) d\nu, \\ \tilde{B}_{nm} &= \iint_S V(\theta, \lambda) \tilde{Y}_{nm}^s(\theta, \lambda) d\nu.\end{aligned}\quad (20.47)$$

The more complicated formulae for ordinary spherical harmonic functions differ from these only in so far as they are divided by the appropriate norms squared (38).

Since there is an implied relationship between the density  $\sigma$  and the potential coefficients  $A_{nm}$ ,  $B_{nm}$ , one is tempted to have a closer look at this relation. For instance, can knowledge of the potential coefficients teach us something about the internal distribution of masses? To answer this question, let us first discuss the (direct) relation of the coefficients with density. Equation (6.25) can be written as

$$W_g(\vec{r}_A) = G \iiint_B \frac{\sigma(\vec{r})}{\rho(\vec{r}_A, \vec{r})} d\mathcal{B}. \quad (20.48)$$

By denoting the spatial angle between  $\vec{r}_A$  and  $\vec{r}$  by  $\psi$ , the inverse of the distance  $\rho$

between the point of interest  $A$  and the dummy point  $\bar{r}$  can be expressed as (cf. (8.3) and (8.4))

$$\rho^{-1}(\bar{r}_A, \bar{r}) = (r_A^2 + r^2 - 2rr_A \cos \psi)^{-1/2} = \frac{1}{r_A} \sum_{n=0}^{\infty} \left( \frac{r}{r_A} \right)^n P_n(\cos \psi). \quad (20.49)$$

Expressing  $\cos \psi$  by means of spherical coordinates  $\theta_A, \lambda_A$  of  $A$  and spherical coordinates  $\theta, \lambda$  of the dummy point, from spherical trigonometry we get

$$\cos \psi = \cos \theta_A \cos \theta + \sin \theta_A \sin \theta \cos(\lambda - \lambda_A). \quad (20.50)$$

Through tedious computations, it can then be proved that the substitution for  $\cos \psi$  from (50) to (49) yields [HOBSON, 1931]:

$$\begin{aligned} P_n(\cos \psi) &= P_n(\cos \theta_A) P_n(\cos \theta) \\ &+ 2 \sum_{m=1}^n \frac{(n-m)!}{(n+m)!} (Y_{nm}^c(\theta_A, \lambda_A) Y_{nm}^c(\theta, \lambda) \\ &+ Y_{nm}^s(\theta_A, \lambda_A) Y_{nm}^s(\theta, \lambda)). \end{aligned} \quad (20.51)$$

This formula was already known to Legendre, and here it will be called the *Legendre decomposition formula*. Substituting this result back into (48) and interchanging the summations with volume integration, we end up with a double series of the same shape as that in (44). This is left to the reader to work out. Equating these two series term by term, we finally obtain

$$\begin{aligned} A_{n0} &= \frac{G}{a} \iiint_{\mathcal{B}} \sigma(\bar{r}) \left( \frac{r}{a} \right)^n Y_{n0}^c(\theta) d\mathcal{B}, \\ \left\{ \begin{array}{l} A_{nm} \\ B_{nm} \end{array} \right\} &= \frac{2G}{a} \frac{(n-m)!}{(n+m)!} \iiint_{\mathcal{B}} \sigma(\bar{r}) \left( \frac{r}{a} \right)^n \left\{ \begin{array}{l} Y_{nm}^c(\theta, \lambda) \\ Y_{nm}^s(\theta, \lambda) \end{array} \right\} d\mathcal{B}. \end{aligned} \quad (20.52)$$

These are the equations relating the potential coefficients, in the units of potential, directly to the density distribution  $\sigma$ .

The inverse problem, i.e., the solution for  $\sigma(\bar{r}_A)$  as a function of the potential coefficients, is not possible. However, at least some indirect insight into the density distribution can be gained from the potential coefficients: namely, there is a relation between some potential coefficients and the coordinates of the earth's centre of mass, and the principal moments as well as products of inertia. To derive this relation, let us adopt an arbitrary Cartesian system fixed to the earth. Transforming the spherical coordinates  $r, \theta, \lambda$  of the dummy point used in (52) into these Cartesian coordinates  $\bar{r}' \equiv (x', y', z')$ , the reader can verify that the first twelve spherical

harmonic functions become

$$\begin{aligned}
 Y_{0,0}^c(x', y', z') &= 1, & Y_{0,0}^s(x', y', z') &= 0, \\
 Y_{1,0}^c(x', y', z') &= \frac{z'}{r'}, & Y_{1,0}^s(x', y', z') &= 0, \\
 Y_{1,1}^c(x', y', z') &= \frac{x'}{r'}, & Y_{1,1}^s(x', y', z') &= \frac{y'}{r'}, \\
 Y_{2,0}^c(x', y', z') &= \frac{-x'^2 - y'^2 + 2z'^2}{2r'^2}, & Y_{2,0}^s(x', y', z') &= 0, \\
 Y_{2,1}^c(x', y', z') &= \frac{3x'z'}{r'^2}, & Y_{2,1}^s(x', y', z') &= \frac{3y'z'}{r'^2}, \\
 Y_{2,2}^c(x', y', z') &= \frac{3(x'^2 - y'^2)}{r'^2}, & Y_{2,2}^s(x', y', z') &= \frac{6x'y'}{r'^2}.
 \end{aligned} \tag{20.53}$$

Substituting these back into (52), we discover that the first twelve potential coefficients acquire the following form:

$$\begin{aligned}
 A_{0,0} &= \frac{G}{a} M, & B_{0,0} &= 0, \\
 A_{1,0} &= \frac{G}{a^2} Mz'_c, & B_{1,0} &= 0, \\
 A_{1,1} &= \frac{G}{a^2} Mx'_c, & B_{1,1} &= \frac{G}{a^2} My'_c, \\
 A_{2,0} &= \frac{G}{a^3} \left( \frac{I'_x + I'_y}{2} - I'_z \right), & B_{2,0} &= 0, \\
 A_{2,1} &= \frac{G}{a^3} I'_{xz}, & B_{2,1} &= \frac{G}{a^3} I'_{yz}, \\
 A_{2,2} &= \frac{G}{4a^3} (I'_y - I'_x), & B_{2,2} &= \frac{G}{2a^3} I'_{xy},
 \end{aligned} \tag{20.54}$$

all in the units of potential. Here,  $x'_c, y'_c, z'_c$  are the coordinates of the earth's centre of mass, e.g.,  $x'_c = M^{-1} \iiint_B x' \sigma(\bar{r}) d\mathcal{B}$ , in the adopted Cartesian coordinate system (see FIG. 8). The moments of inertia of the earth are  $I'_x, I'_y, I'_z$ , e.g.,  $I'_x = \iiint_B (y'^2 + z'^2) \sigma(\bar{r}) d\mathcal{B}$ , with respect to the coordinate axes of the primed coordinate system, and  $I'_{xy}, I'_{xz}, I'_{yz}$  are the products of inertia of the earth, e.g.,  $I'_{xy} = \iiint_B x' y' \sigma(\bar{r}) d\mathcal{B}$ , with respect to the same coordinate system [MACMILLAN, 1936]. Evidently, these lower order potential coefficients show not only the location of the earth's centre of mass but also the orientation and dimensions of the ellipsoid of inertia at the origin of the adopted coordinate system. The higher order potential coefficients have a similar but more complex (and thus less useful) dynamic interpretation.

The relations (54) are, therefore, of profound importance when we wish to realize the natural geocentric coordinate system introduced in §5.3. If the coordinate system is selected so as to make  $A_{1,0} = A_{1,1} = B_{1,1} = 0$ , then the system is geocentric. If, in

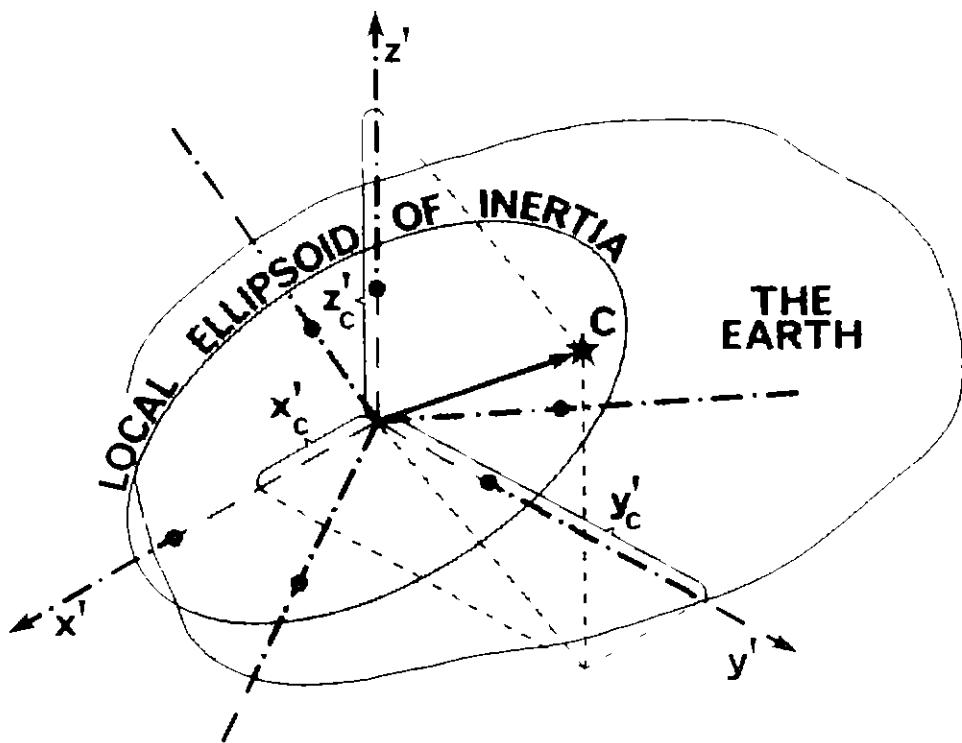


FIG. 20.8. Local ellipsoid of inertia.

addition,  $A_{2,1} = B_{2,1} = B_{2,2} = 0$ , then the system becomes coaxial with the principal ellipsoid of inertia, i.e., it becomes the natural geocentric system. In this case,  $I'_x = I_1$ ,  $I'_y = I_2$ ,  $I'_z = I_3$ , and we have

$$A_{2,0} = \frac{G}{a^3} \left( \frac{I_1 + I_2}{2} - I_3 \right), \quad A_{2,2} = \frac{G}{4a^3} (I_2 - I_1). \quad (20.55)$$

This new (natural geocentric) system is related to the originally adopted system through the following transformation equations:

$$\bar{r} = \mathbf{R}(\omega_1, \omega_2, \omega_3)(\bar{r}' - \bar{r}_c'), \quad (20.56)$$

where  $\bar{r}' = (x'_c, y'_c, z'_c)$ , and  $\mathbf{R}(\omega_1, \omega_2, \omega_3)$  is the rotation matrix (cf. §3.3). The three rotation angles are obtained by the eigenvalue diagonalization (see §3.1) of the principal (i.e., evaluated for the centre of mass of the earth) tensor of inertia:

$$\begin{bmatrix} I_x & I_{xy} & I_{xz} \\ I_{yz} & I_y & I_{yz} \\ I_{zx} & I_{zy} & I_z \end{bmatrix} \rightarrow \begin{bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{bmatrix}. \quad (20.57)$$

As already seen, the potential coefficients are all in the physical units of potential, e.g.,  $\text{cm}^2 \text{s}^{-2}$ . It is often convenient to work with unitless potential coefficients. For

this purpose, the gravitational potential (44) is often written as

$$\boxed{W_g(r; \theta, \lambda) = \frac{GM}{r} \left[ 1 - \sum_{n=1}^{\infty} \left( \frac{a}{r} \right)^n \sum_{m=0}^n (J_{nm} \cos m\lambda + K_{nm} \sin m\lambda) P_{nm}(\cos \theta) \right], \quad (20.58)}$$

where, evidently,  $J_{nm} = -A_{nm}a/(GM)$ , and  $K_{nm} = -B_{nm}a/(GM)$ . Note that when all the potential coefficients are neglected, one gets the spherical approximation of the gravitational potential:

$$W_g^S(r) = \frac{GM}{r}, \quad (20.59)$$

which corresponds to (4). Observations have shown that  $J_{2,0} \doteq 1082.63 10^{-6}$ , much larger than the rest of the potential coefficients which are, at most, of the order of  $10^{-6}$  [IUGG, 1976]. It will be seen in the next section that the  $J_2 = J_{2,0}$  term reflects the ellipticity of the earth, while the rest of the coefficients reflect the remaining irregularities. In the natural coordinate system, defined by (56), we have

$$J_{2,0} = J_2 = -\frac{\left(\frac{1}{2}(I_1 + I_2) - I_3\right)}{Ma^2} \doteq \frac{I_3 - I_1}{Ma^2} \quad (20.60)$$

(cf. §5.3). Hence the elliptical approximation of the gravitational potential reads

$$\begin{aligned} W_g^E(r, \theta) &\doteq \frac{GM}{r} - \frac{G(I_3 - I_1)}{r^3} P_2(\cos \theta) \\ &= \frac{GM}{r} \left( 1 + \frac{I_3 - I_1}{2Mr^2} \right) - \frac{3G(I_3 - I_1)}{2r^3} \cos^2 \theta. \end{aligned} \quad (20.61)$$

So far, our attention has been directed solely to the development of the gravitational potential into spherical harmonics. A very similar argument can be put forward to arrive at the development of  $W_g$  into *ellipsoidal harmonic functions*. This will not be done here; we will content ourselves with quoting the following result from HOBSON [1931]:

$$\boxed{W_g(u, \theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=0}^n q_{nm}(u, E, b) (A_{nm} Y_{nm}^c(\Theta, \lambda) + B_{nm} Y_{nm}^s(\Theta, \lambda)), \quad (20.62)}$$

where  $q_{nm}(u, E, b) = Q_{nm}(iu/E)/Q_{nm}(ib/E)$ ,  $i = \sqrt{-1}$ , and  $Q_{nm}$  are *Legendre functions of the second kind* [ABRAMOWITZ AND STEGUN, 1964]. This equation is valid for the outside of a boundary ellipsoid  $(E, b)$ . The potential coefficients are evaluated from formulae parallel to (47). Here, the ellipsoid  $(E, b)$  plays the same role as the sphere  $r = a$  played in the spherical case. Also note that both series (44) and (62)

have a very similar structure. The main difference is in the radial terms where the ellipsoidal variety are much more complicated.

Let us close this section by restating that an attempt to even set up the boundary value problem for  $W_g$  has not yet been made. The case dealt with so far was purely hypothetical, since we have no means of providing the values of the gravitational potential for the evaluation of potential coefficients on either the boundary sphere or the boundary ellipsoid. The way of evaluating the potential coefficients from observable quantities will be discussed in Chapters 22 and 23.

### 20.3. Model gravity field

In a variety of tasks, it is advantageous to work with a model gravity field, as was done in §6.2 and §17.4. The degree to which this model field should approximate the actual gravity field depends on the task for which it serves. The simplest model is the *radial field*. This can be thought of as being generated by either a particle of a negligible size and mass comparable to that of the earth, or by a sphere with stratified distribution of masses which would, of course, produce an identical model field outside the sphere. The potential of such a field is given by (59), which shows that the field is a function only of the distance from the centre of the field. Its equipotential surfaces are concentric, spherical surfaces. This model has already been used in §6.2 to determine an approximate value for the vertical gradient of gravity. It is also used extensively when working with satellites, as will be seen in Chapter 23.

A closer approximation to reality is an *ellipsoidal model field*. This model  $U(\phi, \lambda, h)$  has already been seen in §6.2, now details will be provided. In geodesy, it is customary to seek a model field that can be thought of as

(a) sharing with the actual field—and thus with the earth—the spin angular velocity  $\omega$ ;

(b) being generated by the best-fitting geocentric biaxial ellipsoid (cf. §7.3) defined by  $a$  and  $b$ ;

(c) having one of its equipotential surfaces—that on which the potential  $U_0$  is equal to the potential  $W_0$  of actual gravity on the geoid—coincident with the ellipsoidal surface.

A model with such properties is called the normal gravity field and its potential is denoted simply by  $U$  (cf. §6.2).

Of course, such a normal field is obtainable only to a certain level of approximation that reflects the present level of knowledge about the actual gravity field. Hence, necessarily, there does not exist the perfect normal gravity field, and, consequently, the existing normal fields (cf. §6.2, §7.3) must be regarded as only approximations of the theoretical ideal. Also, it must be remembered that the definition of a normal field neither requires nor stipulates the knowledge of unique mass distribution within the generating ellipsoid. The situation is thus similar to that of the radial field discussed above, where different mass distributions all produce one and the same field. We shall now show that the above requirements define a unique normal field.

To prove this, it is advantageous to work with the EL coordinate system introduced in §20.1. To enforce the first requirement, (a), the normal potential  $U(u, \Theta)$  must be expressed as a sum of  $W_c(u, \Theta)$ —the centrifugal potential—and  $W_g^N$  which is the part of  $W_g(u, \Theta)$  that is needed to satisfy the other two requirements. Thus

$$U(u, \Theta) = W_g^N(u, \Theta) + W_c(u, \Theta). \quad (20.63)$$

The potential  $W_c(u, \Theta)$  can be easily formulated by referring to FIG. 9:

$$W_c(u, \Theta) = \frac{1}{2}\omega^2(u^2 + E^2)\sin^2\Theta. \quad (20.64)$$

According to the second, (b), and third, (c), requirements, both the normal gravity potential  $U$  and the normal gravitational potential  $W_g^N$  are symmetrical, i.e., they are not functions of  $\lambda$ .

The third requirement, (c), stipulates that

$$U(b, \Theta) = W_0. \quad (20.65)$$

This can be regarded as the equation of a geocentric ellipsoid given by  $b$  and  $E$  (or any two equivalent parameters). Substituting for  $U$  from (63) and (64), we can rewrite (65) as

$$W_g^N(b, \Theta) + \frac{1}{2}\omega^2a^2\sin^2\Theta = W_0. \quad (20.66)$$

All that has to be done now is to find a  $W_g^N(b, \Theta)$  that satisfies this equation.

Since  $W_g^N$  must be symmetrical, any series expression, such as (62), may contain only terms with  $m=0$ , i.e., the zonal terms. Therefore, the general expression for

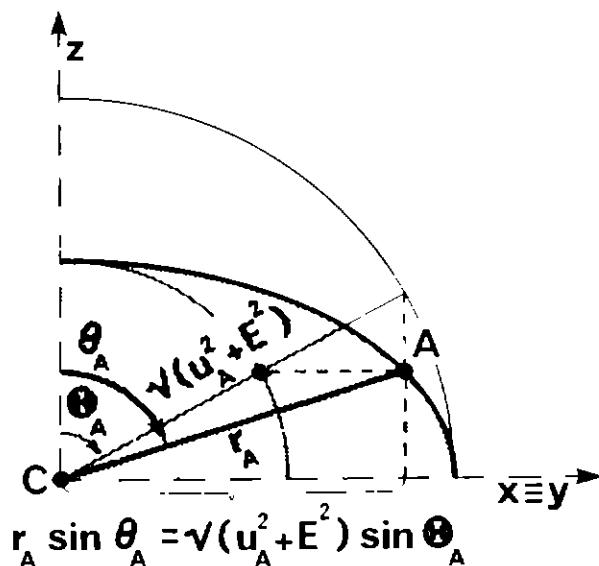


FIG. 20.9. Centrifugal potential in ellipsoidal coordinates.

$W_g^N$  in ellipsoidal coordinates will be

$$W_g^N(u, \Theta) = \sum_{n=0}^{\infty} q_n(u, E, b) A_n P_n(\cos \Theta). \quad (20.67)$$

But on the ellipsoid ( $b, E$ ),  $u = b$  and evidently all the radial terms are equal to one, because

$$q_n(b, E, b) = Q_{nm}\left(i \frac{b}{E}\right) / Q_{nm}\left(i \frac{b}{E}\right) = 1. \quad (20.68)$$

So the equation of the ellipsoid (66) becomes

$$\begin{aligned} \sum_{n=0}^{\infty} A_n P_n(\cos \Theta) &= W_0 - \frac{1}{2} \omega^2 a^2 \sin^2 \Theta \\ &= W_0 P_0(\cos \Theta) - \frac{1}{2} \omega^2 a^2 \frac{2}{3} [P_0(\cos \Theta) - P_2(\cos \Theta)]. \end{aligned} \quad (20.69)$$

The only way this equation can be satisfied for any value of  $\Theta$  is if all the zonal coefficients multiplying the corresponding Legendre functions on the left- and right-hand sides are equal. Thus we have

$$\begin{aligned} A_0 &= W_0 - \frac{\omega^2 a^2}{3}, \quad A_1 = 0, \\ A_2 &= \frac{\omega^2 a^2}{3}, \quad A_n = 0, \quad n = 3, 4, \dots, \end{aligned} \quad (20.70)$$

and the normal gravitational potential becomes

$$W_g^N(u, \Theta) = q_0(u, E, b) \left( W_0 - \frac{\omega^2 a^2}{3} \right) + q_2(u, E, b) \frac{\omega^2 a^2}{3} P_2(\cos \Theta). \quad (20.71)$$

Finally, we want to express the normal gravity potential  $U$  as a function of  $GM$  that can be directly determined, rather than as a function of  $W_0$  that cannot. In other words, an attempt will be made to eliminate  $W_0$  from (71). To do this,  $q_0$  is expressed as a function of  $r$  and

$$q_0(u, E, b) = \arctan\left(\frac{E}{u}\right) / \arctan\left(\frac{E}{b}\right) \doteq \frac{E}{r} \arctan^{-1}\left(\frac{E}{b}\right) \quad (20.72)$$

is obtained to an accuracy of  $(1/r)^3$ . Then comparing  $W_g^N$  with  $W_g$ , both taken to the accuracy of  $(1/r)^3$  only, we get the relation we have been seeking:

$$\frac{E}{r} \arctan^{-1}\left(\frac{E}{b}\right) \left( W_0 - \frac{\omega^2 a^2}{3} \right) \doteq \frac{GM}{r}. \quad (20.73)$$

By expressing  $W_0$  in terms of  $GM$  and substituting back to (66), the normal potential

in ellipsoidal coordinates is finally obtained in the following form:

$$\boxed{U(u, \Theta) = \frac{GM}{E} \arctan\left(\frac{E}{u}\right) + \frac{\omega^2(u^2 + E^2)}{3} (1 - P_2(\cos \Theta)) + q_2(u, E, b) \frac{\omega^2 a^2}{3} P_2(\cos \Theta).} \quad (20.74)$$

Evidently, the normal potential is defined for every point  $(u, \Theta)$  once  $GM$ ,  $\omega$ , and the reference ellipsoid  $(b, E)$  are specified. This concludes the proof that the above requirements ((a) to (c)) uniquely define the normal field.

For many applications, it is convenient to express the normal potential in spherical coordinates. To obtain the appropriate expression, we begin with the development of the gravitational potential  $W_g(r, \theta, \lambda)$  into spherical harmonics (58) and seek its normal part  $W_g^N$ . This part must satisfy the definition equation (cf. (63)),

$$U(r, \theta) = W_g^N(r, \theta) + W_c(r, \theta), \quad (20.75)$$

as well as the requirements ((a) to (c)) spelled out at the beginning of this section. Here, once more, the normal gravitational potential must be symmetrical, and thus only the zonal harmonics have to be considered. Moreover, since the normal field is also symmetrical with respect to the equator, all the harmonics of odd degree ( $n = 2k + 1, k = 0, 1, \dots$ ) must disappear. The normal gravity potential is then sought in the following form (cf. (61)):

$$\boxed{U(r, \theta) = \frac{GM}{r} \left( 1 - \sum_{n=2,4,6,\dots}^{\infty} \left( \frac{a}{r} \right)^n J_n^N P_n(\cos \theta) \right),} \quad (20.76)$$

where the *normal potential coefficients*  $J_n^N = J_{n0}^N$  are functions of all the required parameters, i.e., the size and the shape of the geocentric reference ellipsoid, the earth's spin velocity, and the earth's mass. For instance,  $J_2^N$ , which has the centrifugal term absorbed in it, can be shown to be equal to [DE SITTER, 1924]

$$J_2^N = \frac{2}{3}f - \frac{1}{3}m - \frac{1}{3}f^2 + \frac{2}{21}fm, \quad (20.77)$$

where  $m$  is the geodetic factor defined by eqn. (7.23). For practical purposes, it is sufficient to take just a few of these normal potential coefficients, as was done for the International Geodetic Reference System 1967 [IAG, 1971]. These coefficients, together with  $GM$ ,  $a$ , and  $\omega$  (cf. (76)), then specify uniquely the normal gravity potential and, thus, even the normal gravity field.

It should be pointed out that for the purpose of defining the normal gravity field, the mass of the earth is generally considered to include the mass of the atmosphere, i.e., the mass of the earth is augmented by about  $0.89 \times 10^{-6} \cdot M$  (cf. §9.4). This is done so that one can regard the normal gravity field generated by the ellipsoid as propagating in empty space above the ellipsoid.

Before talking about normal gravity, let us clarify one more point. So far, the radial and ellipsoidal model fields have been considered. Although in the vast majority of cases the ellipsoidal field, the normal in particular, is an adequate model, there are instances when an even better approximation to the actual gravity field is needed, and the mathematical simplicity of the model has to be sacrificed to this end. In such instances, a truncated spherical or ellipsoidal series, whose potential coefficients are determined from the actual field data, can be used as a model (cf. §24.4). Evidently, the equipotential surfaces of such a higher order model field are not as smooth as those of the normal field; on the other hand, they do not depart from the equipotential surfaces of the actual field as much. An example of a higher order model field, the spheroids, was seen in §7.2. It should be noted that such higher order reference fields may be neither axially symmetrical, i.e., they may depend on  $\lambda$ , nor symmetrical with respect to the equator, i.e., they may contain odd order harmonics. For example, a pear-shaped reference surface will contain the (3,0) term (cf. §20.2).

Any model gravity field has model gravity associated with it. It is, of course, defined as the gradient of the model field potential (cf. (6.31)). We have already worked with the gravity generated by a radial field (cf. (6.11)) and have also mentioned a few formulae for normal gravity (cf. (6.15) to (6.19)). Now we are in a position to show how these expressions are derived for normal gravity.

The most convenient coordinate system for this task is the EL system. In seeking the gradient of the normal gravity potential as given in (74), the gradient operation is first expressed in the ellipsoidal system. We obtain

$$\begin{aligned}\bar{\gamma}(u, \Theta) = \nabla U(u, \Theta) &= \sqrt{\frac{u^2 + E^2}{u^2 + E^2 \cos^2 \Theta}} \frac{\partial U}{\partial u} \bar{e}_u \\ &+ \frac{1}{\sqrt{u^2 + E^2 \cos^2 \Theta}} \frac{\partial U}{\partial \Theta} \bar{e}_{\Theta} + \frac{1}{\sqrt{(u^2 + E^2) \sin \Theta}} \frac{\partial U}{\partial \lambda} \bar{e}_{\lambda}.\end{aligned}\quad (20.78)$$

The third term on the right-hand side must disappear since  $U$  is not a function of  $\lambda$ .

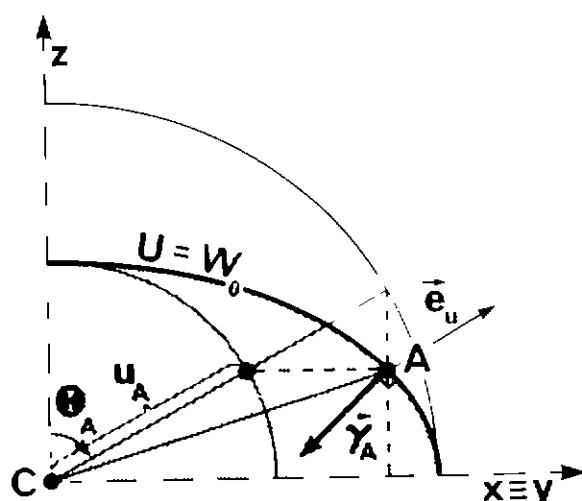


FIG. 20.10. Direction of normal gravity.

Also, in the vicinity of the geocentric reference ellipsoid, normal gravity is directed approximately along the direction of  $u$  when the flattening of the ellipsoid is small, as can be seen in FIG. 10. Nowhere is the deviation of these two directions greater than some 13 minutes of arc. Thus, to a high degree of accuracy (better than 0.2  $\mu\text{Gal}$ ) even the second term containing the rate of change with  $\Theta$  can be dropped, and we can write, for the magnitude of normal gravity (the negative sign is because of the opposite directions of  $\vec{e}_u$  and  $\vec{\gamma}$ ),

$$\gamma(u, \Theta) \doteq -\sqrt{\frac{u^2 + E^2}{u^2 + E^2 \cos^2 \Theta}} \frac{\partial U}{\partial u}. \quad (20.79)$$

Evaluating now the partial derivative of  $U$  from (74), where the derivative of  $q_2(u, b, E)$  is equal to  $-3b^3/u^4$  (with a relative accuracy better than  $e^4$ ), we obtain

$$\begin{aligned} \gamma(u, \Theta) \doteq & -\sqrt{\frac{u^2 + E^2}{u^2 + E^2 \cos^2 \Theta}} \left[ -\frac{GM}{u^2 + E^2} - \frac{a^2 \omega^2 b^3}{u^4} P_2(\cos \Theta) \right. \\ & \left. + \frac{2\omega^2 u}{3} (1 - P_2(\cos \Theta)) \right]. \end{aligned} \quad (20.80)$$

When working with normal gravity, it is customary to first evaluate it on the geocentric reference ellipsoid and then correct for the effect of the location above the ellipsoid, as was done in §6.2. This is because the vertical gradient of normal gravity can be derived rather easily, as will be seen in the next chapter. Hence, our first task is to work out a formula for normal gravity on the surface of the ellipsoid, i.e., for  $\gamma_0$ . As already known,  $u = b$  on the ellipsoid, and (80) gives, after some elementary mathematical manipulations,

$$\gamma_0(\Theta) = \gamma(b, \Theta) \doteq \frac{GM}{a\sqrt{a^2 \cos^2 \Theta + b^2 \sin^2 \Theta}} \left( 1 - \frac{2}{3}m + \left( \frac{a^2}{b^2}m + \frac{2}{3}m \right) P_2(\cos \Theta) \right), \quad (20.81)$$

where  $m$  is again the geodetic factor.

From (81) we can derive, by elementary operations, Clairaut's theorem (cf. §7.3). What we are after here, however, are the *Somigliana formulae* of which the first reads

$$\gamma_0(\Theta) \doteq \frac{a\gamma_P \cos^2 \Theta + b\gamma_E \sin^2 \Theta}{\sqrt{a^2 \cos^2 \Theta + b^2 \sin^2 \Theta}}. \quad (20.82)$$

This is accurate to the same order of accuracy as (74). The meaning of  $\gamma_P$  and  $\gamma_E$  is the same as in §7.3.

The second Somigliana formula merely uses (25) to express normal gravity as a function of geodetic latitude, instead of the second ellipsoidal coordinate  $\Theta$ :

$$\gamma_0(\phi) \doteq \frac{a\gamma_E \cos^2 \phi + b\gamma_P \sin^2 \phi}{\sqrt{a^2 \cos^2 \phi + b^2 \sin^2 \phi}}. \quad (20.83)$$

To transform this equation into the shape Cassini used for the first international formula (cf. (6.17)),  $\cos^2\phi$  is first expressed as  $1 - \sin^2\phi$ , and geometrical as well as gravity flattenings are used. After a few steps, we get

$$\gamma_0(\phi) \doteq \gamma_E \frac{1 + (\tilde{f} - f - ff) \sin^2\phi}{\sqrt{1 + (f^2 - 2f) \sin^2\phi}}. \quad (20.84)$$

Developing the denominator into a power series and neglecting higher order terms, we finally arrive at the general expression

$$\boxed{\gamma_0(\phi) \doteq \gamma_E (1 + \alpha \sin^2\phi + \beta \sin^2 2\phi)}, \quad (20.85)$$

where  $\alpha \doteq \tilde{f}$  is the gravity flattening (§7.3), and  $\beta \doteq f(f - \tilde{f})/4$ . This formula is accurate only to the order of  $e^2$ , i.e., to about 50  $\mu\text{Gal}$ . Higher order approximations may be found in, e.g., IAG [1971].

To complete this section, let us state that similar formulae for normal gravity can also be derived from (76) using spherical instead of ellipsoidal coordinates. LEVALLOIS [1970], for instance, gives expressions for  $\alpha, \beta$  as functions of normal potential coefficients derived in this way.

## 20.4. Disturbing potential

The main application of the normal gravity field is in obtaining anomalous or *disturbing potential*  $T$  which is defined as

$$T(\bar{r}_A) = W(\bar{r}_A) - U(\bar{r}_A). \quad (20.86)$$

The quantity  $T$  depicts the regional and local irregularities of  $W$ . Since  $U$  models the bulk of the actual gravity field  $W$ , the disturbing potential is much smaller than either of the other two, and, therefore, any approximation used in evaluating  $T$  is far less critical than approximations used in evaluating the other two potentials.

Because of the definition of the normal gravity field, the disturbing potential satisfies the Laplace equation outside the earth. This can easily be shown by substituting for  $U$  in (86) from (63) and splitting  $W$  into gravitational and centrifugal potentials:

$$\begin{aligned} T(\bar{r}_A) &= W_g(\bar{r}_A) + W_c(\bar{r}_A) - (W_g^N(\bar{r}_A) + W_c(\bar{r}_A)) \\ &= W_g(\bar{r}_A) - W_g^N(\bar{r}_A). \end{aligned} \quad (20.87)$$

Neglecting the atmosphere, (cf. (15)) one immediately gets

$$\boxed{\nabla^2 T(\bar{r}_A) = 0} \quad (20.88)$$

everywhere outside the earth.

For many purposes, it is convenient to seek the disturbing potential in spherical harmonics. Substituting into (86) from (58) and (76), we readily obtain

$$\begin{aligned} T(r, \theta, \lambda) = & \frac{GM}{r} \sum_{n=2,4,\dots}^{\infty} \left(\frac{a}{r}\right)^n (J_n^N - J_n) P_n(\cos \theta) \\ & - \frac{GM}{r} \sum_{n=1,3,\dots}^{\infty} \left(\frac{a}{r}\right)^n J_n P_n(\cos \theta) \\ & - \frac{GM}{r} \sum_{n=1}^{\infty} \left(\frac{a}{r}\right)^n \sum_{m=1}^n (J_{nm} \cos m\lambda + K_{nm} \sin m\lambda) P_{nm}(\cos \theta). \end{aligned} \quad (20.89)$$

This equation is often written as

$$T(r, \theta, \lambda) = \sum_{n=1}^{\infty} T_n(r, \theta, \lambda), \quad (20.90)$$

where  $T_n$  are regarded as being the components of  $T$  for the appropriate orders  $n$ , e.g.,

$$T_5(r, \theta, \lambda) = -\frac{GM}{r} \left(\frac{a}{r}\right)^5 \sum_{m=0}^5 (J_{5m} \cos m\lambda + K_{5m} \sin m\lambda) P_{5m}(\cos \theta).$$

Note that for the properly selected normal field,  $J_2^N$  is equal to  $J_2$ . Also, if the spherical coordinate system coincides with the natural geocentric system, some potential coefficients disappear (cf. §20.2) and we get, in particular,

$$\begin{aligned} T(r, \theta, \lambda) = & \frac{GMa^2}{r^3} J_{2,2} \cos 2\lambda P_{2,2}(\cos \theta) \\ & + \frac{GM}{r} \sum_{n=4,6,\dots}^{\infty} \left(\frac{a}{r}\right)^n (J_n^N - J_n) P_n(\cos \theta) \\ & - \frac{GM}{r} \sum_{n=3,5,\dots}^{\infty} \left(\frac{a}{r}\right)^n J_n P_n(\cos \theta) \\ & - \frac{GM}{r} \sum_{n=3}^{\infty} \left(\frac{a}{r}\right)^n \sum_{m=1}^n (J_{nm} \cos m\lambda + K_{nm} \sin m\lambda) P_{nm}(\cos \theta). \end{aligned} \quad (20.91)$$

This can be written as

$$T(r, \theta, \lambda) = \sum_{n=2}^{\infty} T_n(r, \theta, \lambda), \quad (20.92)$$

with the understanding that  $T_2$  contains only the  $J_{2,2}$  term which is, in any case, very small since, as we saw in §5.3,  $I_1 \doteq I_2$  (cf. (55)).

It is interesting at this stage to also have a look at what happens to the disturbing potential if the mass of the geocentric ellipsoid generating the normal field—denoted

here by  $M^N$ —is assessed improperly. In this case, (90) and (92) will have an additional absolute term:

$$\delta T = -\delta U = T_0 = \frac{GM}{r} - \frac{GM^N}{r} = -\frac{G}{r}(M^N - M) = -\frac{G}{r}\delta M. \quad (20.93)$$

As our present knowledge of the accuracy of the value of  $GM$  is better than  $10^{-6}GM$  [IUGG, 1976], the corresponding error  $T_0$  would cause a systematic (constant) global error in gravity anomalies of the order of 1 mGal at most. Comparable relative errors (of the order of  $10^{-6}$ ) in other parameters of the normal field, i.e., in  $\omega$ ,  $a$ , and  $J_2$ , have a much smaller effect on  $T$ , as one can see from (89).

Let us now turn to an alternative formulation for the disturbing potential based on the integral equation (17). For simplicity, however, we shall consider only the formulation for the surface  $S$  of the earth. To arrive at the equation we are seeking, let us apply Green's third identity to the normal part  $W_g^N$  of  $W_g$ . We get

$$\oint_S \left( \frac{1}{\rho} \frac{\partial W_g^N}{\partial n} - W_g^N \frac{\partial \rho^{-1}}{\partial n} \right) dS - \iiint_B \frac{1}{\rho} \nabla^2 W_g^N d\mathcal{B} = 2\pi W_g^N(\bar{r}_A),$$

for  $A$  on  $S$ , (20.94)

where  $n$  is, again, the outer normal to the earth's surface.

Note that in order to be able to write the equation in this form, the normal gravitational potential  $W_g^N$  must be given a different meaning while preserving its analytical form as developed in §20.3. We have to assume that, this time, the normal gravitational potential is generated by the earth itself rather than by the geocentric ellipsoid. Thus, a model earth must be assumed to exist which is identical with the real earth in both size and shape but different in the distribution of masses. This is perfectly admissible when one realizes that the normal gravitational field can be treated as a purely conventional quantity without even having to stipulate the generating body at all. Again, we can write (cf. (12))

$$\nabla^2 W_g^N(\bar{r}) = -4\pi G \sigma^N(\bar{r}), \quad (20.95)$$

where  $\sigma^N(\bar{r})$  is the mass density within the earth that would generate the required normal gravitational potential. Substituting this into (94), we obtain

$$\oint_S \left( \frac{1}{\rho} \frac{\partial W_g^N}{\partial n} - W_g^N \frac{\partial \rho^{-1}}{\partial n} \right) dS = -2\pi W_g^N(\bar{r}_A), \quad (20.96)$$

when realizing that

$$G \iiint_B \frac{\sigma^N(\bar{r})}{\rho(\bar{r}, \bar{r}_A)} d\mathcal{B} = W_g^N(\bar{r}_A). \quad (20.97)$$

After subtracting (96) from (18) and taking (87) into consideration, we get, following a simple rearrangement,

$$T(\bar{r}_A) - \frac{1}{2\pi} \oint_S \left[ T(\bar{r}) \frac{\partial \rho^{-1}}{\partial n} - \frac{1}{\rho} \frac{\partial T(\bar{r})}{\partial n} \right] dS = 0. \quad (20.98)$$

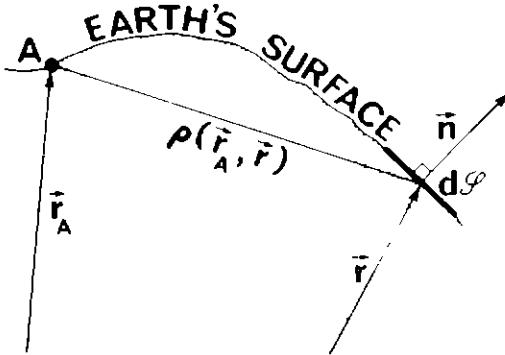


FIG. 20.11. Disturbing potential on the surface of the earth.

This is known as the *Molodenskij integral equation for disturbing potential*  $T$  [MOLODENSKIJ ET AL., 1960]; it is valid on the surface of the earth and is rigorous. The meaning of the individual symbols should be clear from FIG. 11.

In some geodetic applications, it is useful to express the disturbing potential through another physical idealization—an infinitely thin layer of finite *surface density*. To understand this concept, let us first consider a shell  $\mathcal{E}$  of a finite uniform thickness wrapped around an empty earth. It is possible to imagine such a distribution of density  $\sigma^{\mathcal{E}}$  within the shell that, from the outside, its potential will be identical with that of the actual earth [KOCHE AND MORRISON, 1970; OFFICER, 1974]. Expressed mathematically, the gravity potential is

$$W(\vec{r}_A) = G \iiint_{\mathcal{E}} \frac{\sigma^{\mathcal{E}}(\vec{r})}{\rho(\vec{r}, \vec{r}_A)} d\mathcal{E}. \quad (20.99)$$

Next, the thickness of the shell  $\mathcal{E}$  is allowed to become infinitesimally small. At the same time, the density  $\sigma^{\mathcal{E}}$  is changed to  $\sigma^{\mathcal{S}}$  so as to preserve the same potential  $W$ . The expression for  $W$  then becomes

$$W(\vec{r}_A) = G \iint_{\mathcal{S}} \frac{\sigma^{\mathcal{S}}(\vec{r})}{\rho(\vec{r}, \vec{r}_A)} d\mathcal{S}, \quad (20.100)$$

where the integration is carried out over the surface  $\mathcal{S}$  of the earth. It is interesting to realize that while the physical units of  $\sigma^{\mathcal{E}}$  are  $\text{g cm}^{-3}$ , the physical units of  $\sigma^{\mathcal{S}}$  are  $\text{g cm}^{-2}$ ; thus  $\sigma^{\mathcal{E}}$  and  $\sigma^{\mathcal{S}}$  are two different physical quantities. This accounts for the fact that  $\sigma^{\mathcal{S}}$  is finite and not infinite as our intuition would otherwise lead us to believe: the surface density  $\sigma^{\mathcal{S}}$  is only a physical abstraction. It can be shown that there is a unique one to one relation between  $W$  and  $\sigma^{\mathcal{S}}$  [MORRISON, 1972] when the surface  $\mathcal{S}$  is sufficiently smooth.

A similar formulation can also be used for the disturbing potential. Since the disturbing potential  $T$  is much smaller than  $W$ , then even the surface density necessary to generate  $T$  is much smaller than  $\sigma^{\mathcal{S}}$  in (100). Within this context, it is expedient to define a new quantity: the *surface density function*  $\Phi(\vec{r})$ , which is the product of surface density with the gravitational constant [PICK ET AL., 1973]. By

using the surface density function, the disturbing potential can be written as

$$T(\bar{r}_A) = \iint_S \frac{\Phi(\bar{r})}{\rho(\bar{r}, \bar{r}_A)} dS. \quad (20.101)$$

Note that the physical units of  $\Phi$  are those of acceleration, i.e., centimetres per second squared. While the surface density is a useful mathematical tool, its physical interpretation must be treated very carefully particularly in the immediate vicinity of the surface.

The last representation of the disturbing potential to be discussed here is that using ‘buried particles’ of conveniently selected mass called *mascons*. To explain the concept of mascons, let us begin by contemplating the normal potential  $U$  together with the potentials  $P_1, P_2$  of two buried particles of positive and negative masses  $m_1, m_2$ . The potential resulting from the superposition of these three fields is shown in FIG. 12. Evidently, the location and mass of an infinite series of such particles can be specified so as to make their combined potential  $\sum_{i=1}^{\infty} P_i$  equal to the actual disturbing potential  $T$ .

Mathematically, this can be proved in the following way: the gravitational potential  $P_i$  of each mascon  $m_i$  as sensed at  $A$  is given as (cf. (6.25))

$$P_i(\bar{r}_A) = \frac{Gm_i}{\rho}, \quad (20.102)$$

where the meaning of the symbols is clear from FIG. 13;  $m_i$  can be either positive or negative. We can now express  $\rho^{-1}$  in the Legendre functions of  $\cos \psi$  (cf. (3.53)) and develop each Legendre function into a series of spherical harmonics (cf. (51)). Then, summing up  $p$  potentials  $P_i$  and denoting

$$\begin{Bmatrix} A_{nm}^p \\ B_{nm}^p \end{Bmatrix} = \frac{2G}{a} \frac{(n-m)!}{(n+m)!} \sum_{i=1}^p m_i \left( \frac{r_i}{a} \right)^n \begin{Bmatrix} Y_{nm}^c(\theta_i, \lambda_i) \\ Y_{nm}^s(\theta_i, \lambda_i) \end{Bmatrix}, \quad (20.103)$$

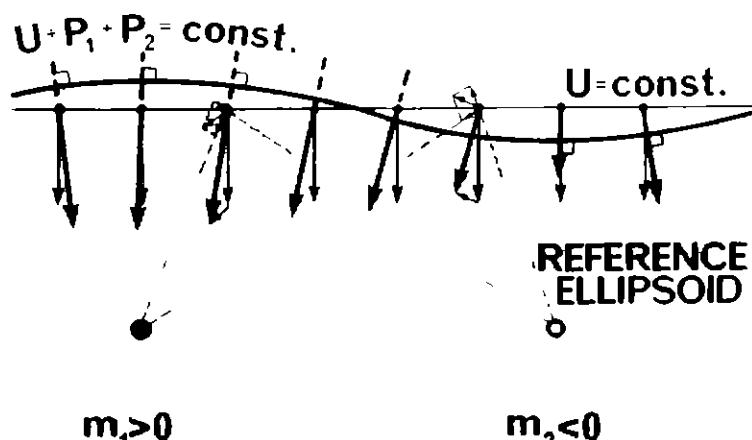
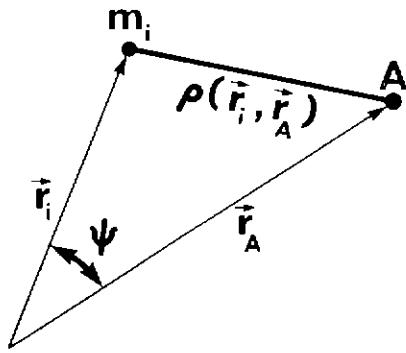


FIG. 20.12. Mascons.



20.13. Potential of a mascon.

we obtain the expression

$$\begin{aligned}
 P^p(r_A, \theta_A, \lambda_A) &= \sum_{i=1}^p P_i(\bar{r}_A) \\
 &= \sum_{n=0}^{\infty} \left( \frac{a}{r_A} \right)^{n+1} \sum_{m=0}^n (A_{nm}^p Y_{nm}^c(\theta_A, \lambda_A) + B_{nm}^p Y_{nm}^s(\theta_A, \lambda_A)).
 \end{aligned} \tag{20.104}$$

Here,  $a$  is selected arbitrarily, and the similarity of (103) to (52) is noted.

It is easily seen that if  $A_{0,0}^p = 0$ , the form of (104) is identical with that of (90) for the disturbing potential  $T$  (realizing that  $B_{0,0}^p = 0$  by definition—cf. (103)). It should be intuitively obvious that  $A_{0,0}^p = 0$  if and only if  $\sum_{i=1}^p m_i = 0$ , i.e., if the amount of positive mass buried in the reference ellipsoid is equal to the amount of negative mass. If this condition is satisfied, these two equations ((90) and (104)) can be equated term by term in  $n$ , and a system of linear equations in  $m_i$  can be obtained. Selecting locations  $\bar{r}_i$  of the buried masses, we can solve for the masses  $m_i$ , or, choosing the masses  $m_i$  beforehand, we can use the equations to solve for the radial distances  $r_i$  of these masses [BALMINO, 1972]. We have to select infinitely many locations  $(r, \theta, \lambda)$  to be able to express the disturbing potential (89) exactly. If only an approximation is needed, then a finite number of mascons can be used.

The advantage of this approach in expressing the disturbing potential is that the mascons can be visualized as point density anomalies within the earth. Then it becomes easier to interpret the disturbing potential  $T$  in terms of density variations, either within the earth's crust or mantle, depending on the depth one uses in placing the mascons.

## CHAPTER 21

### LOCAL TREATMENT OF THE GRAVITY FIELD

Various locally observed parameters of the earth's gravity field usually can be converted into other parameters thus giving a valuable source of information on the field. As well, local features of the earth's gravity field play a certain role in geodetic positioning, as was seen in Part IV. Thus, it is helpful to understand the local behaviour of the gravity field and the local relations between different field parameters.

In this chapter, we begin with the definition of these field parameters and their relation to the (actual) gravity potential. In the second section, the vertical gradient of gravity, with its special place in many geodetic problems, is discussed. Closely related questions of the curvature of the plumb line are dealt with in the third section. Finally, the two dominant causes of local variations of the gravity field, terrain and isostasy, are the content of the last section.

The reader will find the mathematical apparatus used in this chapter markedly different from that of the previous chapter. Also, the notation used here was chosen to suit the local treatment better.

#### 21.1. Conversion of disturbing potential into other field parameters

In Chapter 6, we recognized three basic gravity field parameters that are used in geodesy: the gravity anomaly, the deflection of the vertical, and the geoidal height. We are now going to show the different species of these parameters that can be defined and used.

The gravity anomaly is defined as the scalar whose value is equal to the difference between the magnitude of the actual gravity on the geoid,  $g_0$ , and the normal gravity on the geocentric ellipsoid,  $\gamma_0$  (FIG. 1):

$$\boxed{\Delta g = g_0 - \gamma_0.} \quad (21.1)$$

When  $\Delta g$  is defined in this way, it is called the gravity anomaly on the geoid, or just the *geoidal gravity anomaly*. Often when the meaning is clear from the context, the word 'geoidal' is left out.

It is evident that on land,  $g_0$  has to be generally deduced from the value  $g_A$  observed on the earth's surface. Analogically, gravity observed on the sea bottom has

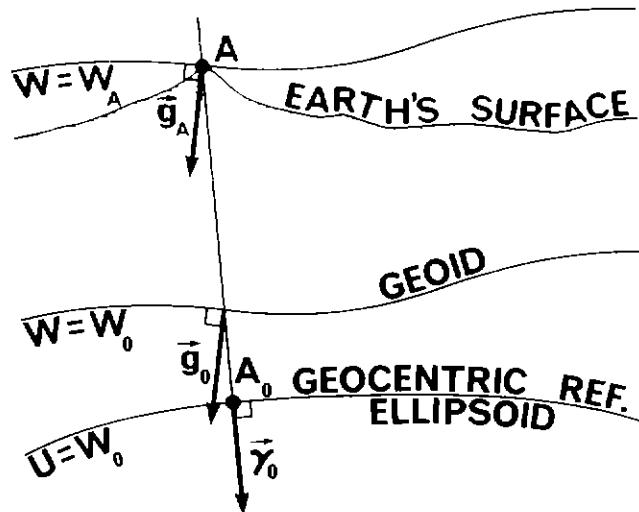


FIG. 21.1. Geoidal gravity anomaly.

to be converted to  $g_0$  by applying a suitable correction. According to the way observed gravity is reduced to the geoid, there are several species of geoidal gravity anomaly of which the free air variety, seen in §6.2, is one. In the next section, the different vertical gravity gradients assumed to be valid between the earth's surface and the geoid will be shown. These lead to the different reduction procedures, and hence different gravity anomalies.

A counterpart to the geoidal anomaly is the *surface gravity anomaly*  $\tilde{\Delta}g$ , which is defined as the difference between the magnitude of the observed gravity taken on the earth's surface and the normal gravity taken on the telluroid (see §7.4):

$$\tilde{\Delta}g = g_A - \gamma_{A'}. \quad (21.2)$$

This kind of gravity anomaly does not require the knowledge of the vertical gradient of the actual gravity within the earth. The exact value of the normal gravity on the telluroid needed for this anomaly can be obtained from, say, (6.15), when the normal height  $H^N$  is used instead of  $h$  (see FIG. 2). Hence, there are no different species of

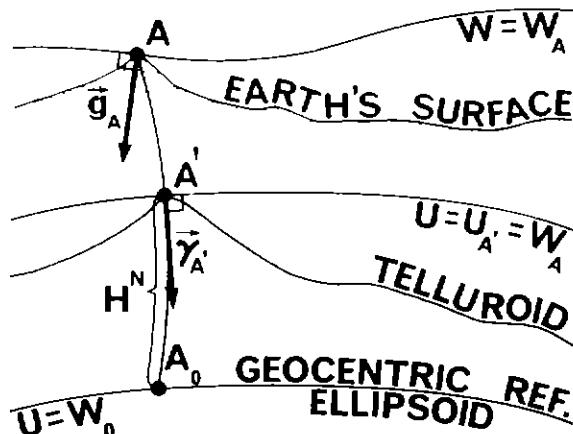


FIG. 21.2. Surface gravity anomaly.

the surface anomaly, as there were of the geoidal anomaly, because the computation of  $\gamma_a$  is unique, and there is no need to reduce the observed gravity  $g_A$ .

A similar situation exists with the deflections of the vertical. In §6.4, the one species of deflection introduced was defined as the spatial angle between the normal gravity vector on the reference ellipsoid and the actual gravity vector on the geoid. This species of  $\theta$  is called the *geoidal deflection*. As shown in FIG. 6.21, it can also be interpreted as the maximum slope of the geoid with respect to the reference ellipsoid at the point of interest.

If the angle is reckoned between the normal gravity vector on the reference ellipsoid and the actual gravity vector on the earth's surface, then we speak about the *surface deflection* and denote it by  $\theta'$  (see FIG. 3). These two deflections differ by an amount dictated by the curvature of the actual plumb line, as already seen in Chapter 6.

The third species of deflection used in geodesy is defined as the angle between the actual gravity vector on the earth's surface and the normal gravity vector on the telluroid (FIG. 4). This angle is called *Molodenskij's deflection* and will be denoted by  $\bar{\theta}$ . Comparing  $\theta'$  with  $\bar{\theta}$  in FIGS. 3 and 4, one quickly comes to the conclusion that the difference between these two is given by the amount of curvature of the plumb line of the normal field, i.e., the curvature of the *normal plumb line* between the reference ellipsoid and the telluroid. More about the curvature of the actual and normal plumb lines will be said in §21.3.

All three of the above deflections are related to the normal-gravity-generating geocentric reference ellipsoid. Their three counterparts, related to the geodetic (i.e., generally non-geocentric) reference ellipsoid, as used in Part IV may also be defined. Chapter 24 will show how these quantities are used in the investigation of the earth's gravity field. Let us just note here that, while the geodetic (ellipsoid related) deflections are useful, it is not customary to define gravity anomalies referred to the geodetic reference ellipsoid. The reason is that the geodetic reference ellipsoid is not thought of as a model figure of the earth and hence there is no normal gravity associated with it.

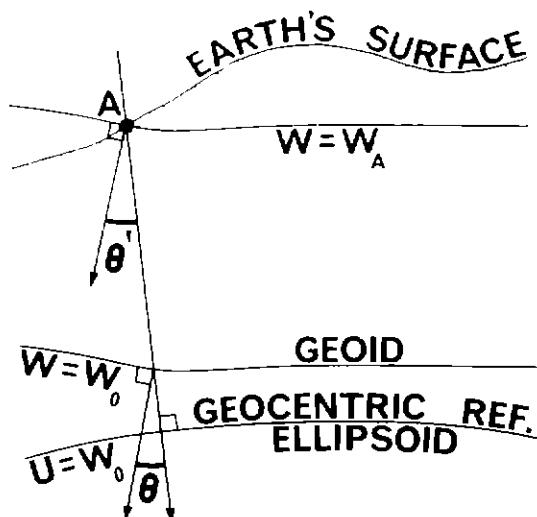


FIG. 21.3. Geoidal and surface deflection of the vertical.

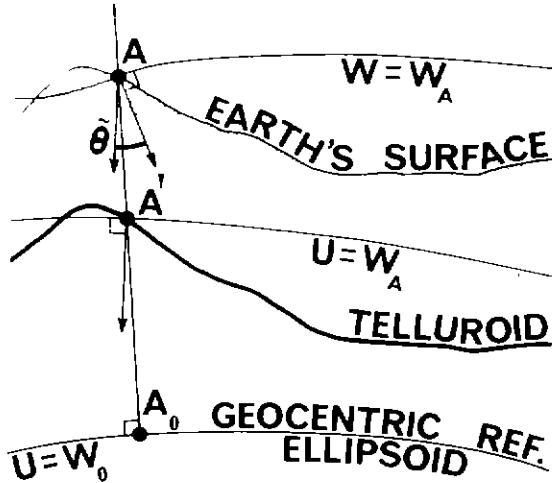


FIG. 21.4. Molodenskij's deflection of the vertical.

The importance of the geoidal height  $N$ , already dealt with extensively in Parts II and IV, has caused us to reintroduce it here. At the same time, the quasigeoidal height  $\xi$ , called the height anomaly (cf. §7.4), also has to be mentioned. In the context of gravity field investigations, the quasigeoid (which is not an equipotential surface and does not have any direct relation to gravity) is often treated as an approximation to the geoid and, as such, has its place in this part of the book.

Let us now have a look at how these quantities are related to the earth's gravity potential and, in particular, to the disturbing potential. To begin with, let us start with the most simple relation, that of the geoidal height. FIG. 5 depicts the situation; assuming the vertical gradient of  $U$  to be constant between the ellipsoid and the geoid, it can be seen immediately that the following equation holds:

$$U_B - U_{B_0} = U_B - W_0 = \frac{\partial U}{\partial n} \Big|_{B_0} N = -\gamma_0 N. \quad (21.3)$$

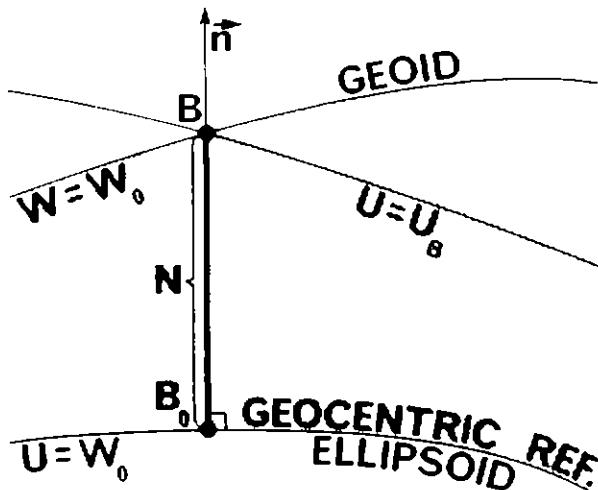


FIG. 21.5. Bruns's formula.

Realizing that  $W_0 - U_B$  is merely the disturbing potential  $T$  at  $B$ , we can write

$$\boxed{N = T/\gamma_0}, \quad (21.4)$$

where the subscript  $B$  has been omitted and it is understood that  $T$  is taken on the geoid. This formula is known as *Brun's formula* and is used extensively in geodesy.

In the above reasoning, it was assumed that the normal gravity field had been properly defined so that the normal potential  $U$  on the ellipsoid was equal to the actual potential on the geoid. As pointed out in §20.4, an error  $\delta M$  in the accepted mass of the earth cannot be ruled out. An error  $\delta M$  in  $M^N$  would cause not only the disturbing potential  $T$  to be in error by  $\delta T$  (cf. (20.93)) but also  $\gamma_0$  would be in error by

$$\delta\gamma_0 = -\frac{G\delta M}{r^2} = \frac{\delta U}{r} = -\frac{\delta T}{r}. \quad (21.5)$$

Then the situation would change from that shown in FIG. 5 to that shown in FIG. 6. Accordingly, (3) would change to

$$U_B - (W_0 - \delta U) = -(\gamma_0 + \delta\gamma_0)(N + \delta N), \quad (21.6)$$

and the *generalized Bruns formula* would read

$$\boxed{N + \delta N = \frac{T + \delta T}{\gamma_0 + \delta\gamma_0}}. \quad (21.7)$$

Here,

$$\delta N = \frac{\delta T}{\gamma_0} \left( 1 + \frac{N}{r} \right) = \frac{\delta T}{\gamma_0}, \quad (21.8)$$

and a simple calculation convinces us that a relative error of  $10^{-6}$  (cf. §20.4) in the

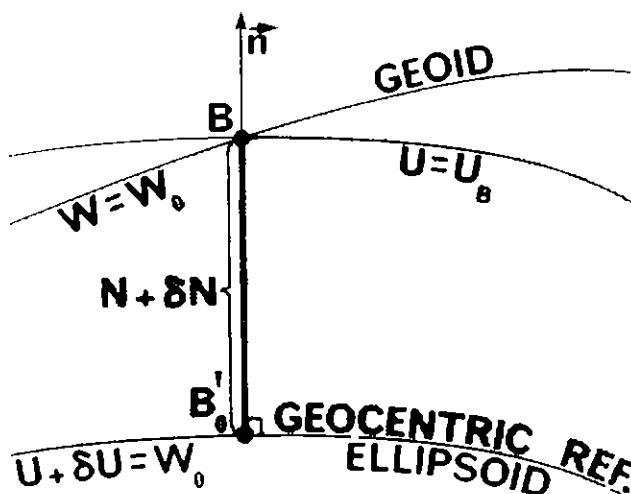


FIG. 21.6. Case of improperly assessed normal potential.

assessment of  $GM$  would show as a constant error of  $|\delta N| \doteq 6$  metres. This must then be interpreted as an error in the size of the implied geocentric reference ellipsoid; this is another illustration of how physical and geometrical quantities are intertwined (cf. §7.3).

Clearly, a formula similar to that of Bruns's can be derived for the height anomaly. Consulting FIG. 7, we can write

$$U_A - W_A = \frac{\partial U}{\partial n} \Big|_{A'} \xi = -\gamma_{A'} \xi. \quad (21.9)$$

Substituting  $T_A$  for the left-hand side, we get

$\xi = T_A / \gamma_{A'},$

(21.10)

where  $T_A$  is taken on the earth's surface and  $\gamma_{A'}$  is on the telluroid. The derivation of the general case follows reasoning similar to that preceding the generalized Bruns formula.

Considering the relation between the geoidal gravity anomaly and the gravity potential  $W$  next, we can write the obvious formula (cf. (1)):

$$\Delta g = |\nabla W|_B - |\nabla U|_{B_0}. \quad (21.11)$$

A more interesting equation, however, is that relating  $\Delta g$  to  $T$ , and it requires a slightly more involved derivation. First, the derivative of (3) is taken in the direction of the ellipsoidal normal  $\vec{n}$ . The derivative of the first term is, to a very high degree of accuracy, equal to  $-\gamma_B$ ; that of the second term is equal to  $-\gamma_0$  exactly. Thus, when it is realized that  $N$  is not a function of  $H$  and thus its derivative along the normal is equal to zero, one obtains (cf. FIG. 8)

$$-\gamma_B + \gamma_0 \doteq -\frac{\partial \gamma}{\partial H} \Big|_{B_0} N. \quad (21.12)$$

Next, the defining equation for  $T$  (20.86) at point  $B$  (on the geoid) is taken, and the

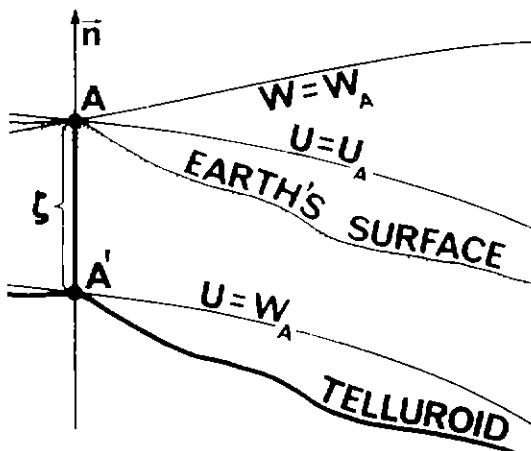


FIG. 21.7. Bruns's formula for height anomaly.

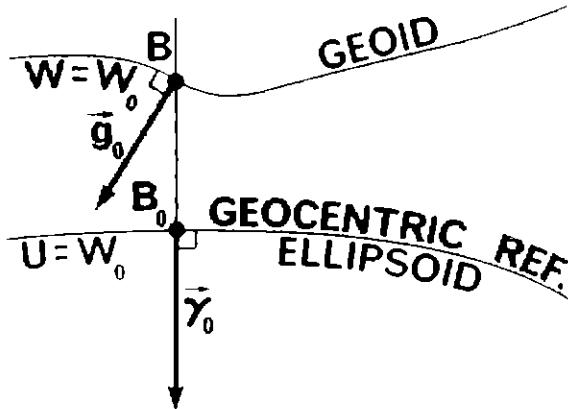


FIG. 21.8. Geoidal gravity anomaly and potential.

derivative is again evaluated with respect to  $H$ . We obtain

$$-g_0 + \gamma_B \doteq \frac{\partial T}{\partial H} \Big|_B. \quad (21.13)$$

Adding (12) and (13), and making use of Bruns's formula, we finally get

$$g_0 - \gamma_0 = \Delta g \doteq -\frac{\partial T}{\partial H} + \frac{1}{\gamma_0} \frac{\partial \gamma}{\partial H} T, \quad (21.14)$$

where both derivatives can be evaluated on the geoid. This formula is known as the *fundamental gravimetric equation* of geodesy. The reader may want to do the derivation to be convinced that if  $M^N$  is not accurately assessed, the above formula reads, to a good enough approximation,

$$\Delta g - \delta \gamma \doteq -\frac{\partial T}{\partial H} + \frac{1}{\gamma_0} \frac{\partial \gamma}{\partial H} (T + \delta T). \quad (21.15)$$

A completely analogous formula holds for the surface gravity anomaly, i.e.,

$$\widetilde{\Delta g} \doteq -\frac{\partial T}{\partial H} + \frac{1}{\gamma} \frac{\partial \gamma}{\partial H} T, \quad (21.16)$$

with the only difference being that the derivatives and  $\gamma$  are evaluated on the telluroid while  $T$  is taken on the surface of the earth [MOLODENSKIJ ET AL., 1960]. Also, the normal for the derivatives is taken with respect to the normal equipotential surface passing through  $A'$  (see FIG. 7).

Finally, the deflection of the vertical  $\theta$  as a function of the earth's gravity potential can be evaluated; we have the obvious formula:

$$\cos \theta = \frac{\bar{\gamma} \cdot \bar{g}}{\gamma \cdot g} = \frac{\bar{\gamma} \nabla W}{\gamma |\nabla W|}, \quad (21.17)$$

which can be used for all three species of deflections. The relation between the

deflection and the disturbing potential can also be derived from (17), but it is clearly more involved. A simpler argument will be followed here.

Let us begin by establishing the relation between the geoidal deflection component  $\xi$ ,  $\eta$  and the geoidal height  $N$ . Since the deflection components at the point of interest can be considered as slopes of the geoid with respect to the reference ellipsoid, in the meridian and prime vertical directions we have (cf. FIG. 9)

$$\xi = -\frac{1}{R} \frac{\partial N}{\partial \phi}, \quad \eta = -\frac{1}{R \cos \phi} \frac{\partial N}{\partial \lambda}, \quad (21.18)$$

where  $R$  is the mean radius of the earth. The negative signs are due to the sign convention for deflection components as normally used in geodesy [BOMFORD, 1971]; the reader should be aware, however, that the sign convention is sometimes reversed for  $\eta$  in North America. Identical formulae link the Molodenskij deflection components with  $\xi$ , when the derivatives are evaluated on the earth's surface. More will be said about this in §22.2. Now we can substitute for  $N$  from Bruns's formula and obtain, realizing that  $\gamma_0$  is not a function of  $\lambda$ ,

$$\xi = \frac{T}{R \gamma_0^2} \frac{\partial \gamma_0}{\partial \phi} - \frac{1}{R \gamma_0} \frac{\partial T}{\partial \phi}, \quad \eta = -\frac{1}{R \gamma_0 \cos \phi} \frac{\partial T}{\partial \lambda}. \quad (21.19)$$

The evaluation of the component of the deflection in a particular azimuth  $\alpha$  can be done from (16.80). It should be intuitively obvious that a constant error  $\delta M$  affecting  $U$  has no effect on the deflection components.

Similar relations for other species of deflections are of no significant use in geodesy and will not be dealt with here. Formulae linking  $\Delta g$ ,  $N$ ,  $\xi$ , and  $\eta$  directly to the surface density distribution, or mascons, instead of the disturbing potential, can be found in the literature (e.g., KOCH AND MORRISON [1970]; BALMINO [1972]).

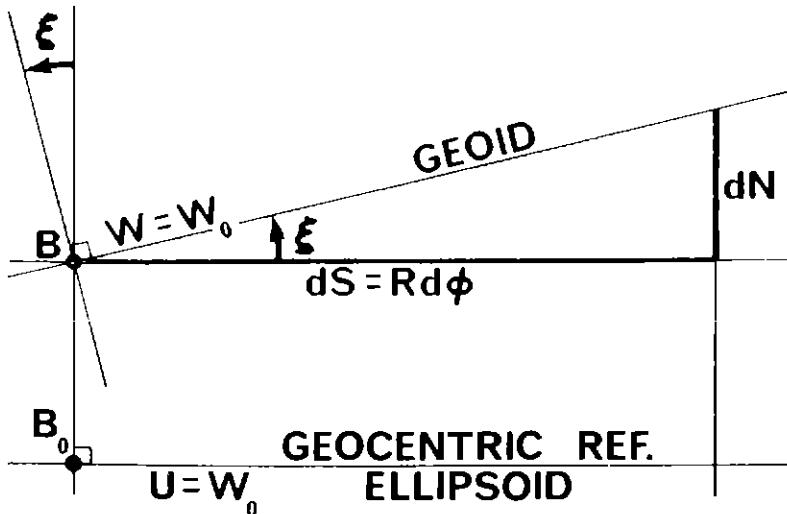


FIG. 21.9. Deflection components and geoidal height.

## 21.2. Vertical gradient of gravity

On several occasions, we have come across the necessity of knowing the vertical gradient of gravity, actual or normal, inside or outside the earth. This topic will now be dealt with systematically.

Let us begin with an investigation of the actual gravity field. Consider a local right-handed Cartesian system of coordinates such that the  $z$ -axis coincides with the outer normal to the equipotential surface  $W = W_A$  (see FIG. 10) at  $A$ . Let the  $x$ -axis point north so that the system is equivalent to the LA system (cf. FIG. 15.3) with the direction of the  $y$ -axis reversed.

The quantity we are looking for is

$$\frac{\partial g}{\partial h} = \frac{\partial g}{\partial H} = \frac{\partial g}{\partial z} = -\frac{\partial^2 W}{\partial z^2}. \quad (21.20)$$

Making use of (20.12), we can rewrite the above as

$$\left. \frac{\partial g}{\partial H} \right|_A = -\left. \frac{\partial^2 W}{\partial z^2} \right|_A = 4\pi G\sigma_A - 2\omega^2 + \left( \left. \frac{\partial^2 W}{\partial x^2} + \frac{\partial^2 W}{\partial y^2} \right) \right|_A, \quad (21.21)$$

where the density  $\sigma_A$  is evaluated according to the location of  $A$  (including fractional density for points on the border of two layers of different densities), and all the terms should be clear except the last. To interpret the last term—the summation of the second derivatives of the potential in horizontal directions—we can use the following trick that involves the equation of the equipotential surface:

$$W(x, y, z) = W_A. \quad (21.22)$$

This (implicit) equation can also be written explicitly as

$$z = z(x, y). \quad (21.23)$$

Following the rules for the total derivatives (see §3.2), the total second derivative of  $W$  with respect, for instance, to  $x$  is

$$\frac{d^2 W}{dx^2} = \frac{\partial^2 W}{\partial x^2} + \frac{\partial^2 W}{\partial x \partial z} \frac{dz}{dx} + \frac{\partial^2 W}{\partial z \partial x} \frac{dz}{dx} + \frac{\partial^2 W}{\partial z^2} \left( \frac{dz}{dx} \right)^2 + \frac{\partial W}{\partial z} \frac{d^2 z}{dx^2}. \quad (21.24)$$

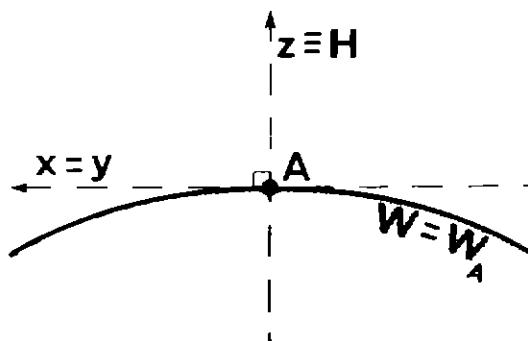


FIG. 21.10. Local Cartesian system (LA system with  $y$ -axis inverted).

But this total derivative is equal to zero because we are dealing with an equipotential surface. Because the equipotential surface has a local extreme at  $A$  in the chosen coordinate system, even  $dz/dx$  becomes zero. Hence, (24) reduces to

$$\frac{\partial^2 W}{\partial x^2} = - \frac{\partial W}{\partial z} \frac{d^2 z}{dx^2} = g \frac{d^2 z}{dx^2}. \quad (21.25)$$

From differential geometry (§3.3), we know that the second derivative  $d^2 z/dx^2$  at  $A$  is equal directly to the curvature  $k_{W_x}$  of the  $z=z(x)$  profile of the surface, since  $z=z(x)$  has an extreme at  $A$ . A similar equation holds for the  $z=z(y)$  profile. By denoting the mean of curvatures  $k_{W_x}$  and  $k_{W_y}$  by  $-J$  (not to be confused with potential coefficients), eqn. (21) becomes

$$\boxed{\frac{\partial g}{\partial H} = -2gJ + 4\pi G\sigma - 2\omega^2.} \quad (21.26)$$

This equation was first formulated by BRUNS [1878] and is valid exactly at any point in space when the proper (fractional) density is associated with points on a boundary between two media of different densities. Note that there is a discontinuity of  $\partial g/\partial H$  at the boundary of layers of different density  $\sigma$ .

Equation (26) can also be applied to the normal gravity field. Limiting ourselves to the space outside the geocentric ellipsoid—where the density is zero—we obtain

$$\frac{\partial \gamma}{\partial H} = -2\gamma J^N - 2\omega^2, \quad (21.27)$$

where  $J^N$  is the mean curvature of the corresponding normal equipotential surface. On the reference ellipsoid, the mean curvature can be evaluated from Euler's formula simply as (see §3.3)

$$J_0^N = \frac{1}{2} \left( \frac{1}{M} + \frac{1}{N} \right), \quad (21.28)$$

where  $M$  and  $N$  are the radii of curvature of the ellipsoid in the meridian and prime vertical directions (cf. §16.2). As was seen earlier, these are functions of the size and shape of the ellipsoid and of latitude  $\phi$ . Substituting for  $M$  and  $N$  in (28), we obtain, after some development [MOLODENSKIY ET AL. 1960],

$$J_0^N \doteq \frac{b}{a^2} (1 + 2f \cos^2 \phi), \quad (21.29)$$

accurate to the order of  $e^2$ . Substituting back in (27) and expressing the corrective term  $2\omega^2$ —much smaller than the other term, cf. §6.1—through the geodetic parameter  $m$  and  $\gamma$ , we finally get the formula for the *normal gravity gradient* in the form used in geodesy:

$$\boxed{\frac{\partial \gamma}{\partial H} \Big|_0 \doteq -\frac{2\gamma_0}{a} (1 + m + 2f \cos^2 \phi).} \quad (21.30)$$

The negative sign, here as well as in (26), shows that the gradient decreases with increasing height, as would be expected. Note that this gradient, already mentioned in §16.4, is used when the normal gravity above the ellipsoid is needed.

Through a few elementary steps, (30) can be transformed to

$$\begin{aligned}\frac{\partial \gamma}{\partial H} \Big|_0 &\doteq -\frac{2\gamma_E}{a} 1.00673(1-0.001415 \sin^2 \phi) \\ &\doteq -0.308745 [\text{mGal/m}] (1-0.001415 \sin^2 \phi).\end{aligned}\quad (21.31)$$

Taking the mean value of  $\sin^2 \phi$  to be 0.4, we get

$$\boxed{\frac{\partial \gamma}{\partial H} \Big|_0 \doteq -0.3086 \text{ mGal/m}.}\quad (21.32)$$

This is the approximate value of the gradient used in the definition of Vignal's heights (see §16.4). The same value is obtained for the *free air gradient* by considering the gravity magnitude to be equal to

$$g \doteq \frac{GM}{r^2} - \omega^2 r \cos^2 \phi.\quad (21.33)$$

Note that this equation differs from (6.11) by the addition of the centrifugal term. Differentiation with respect to  $r$  yields

$$\frac{\partial g}{\partial r} \doteq \frac{\partial g}{\partial H} \doteq -\frac{2GM}{r^3} - \omega^2 \cos^2 \phi.\quad (21.34)$$

Substitution of the value of the mean radius  $R$  of the earth for  $r$  gives

$$\frac{\partial g}{\partial H} \doteq (-0.3083 - 0.000532 \cos^2 \phi) \text{ mGal/m}.\quad (21.35)$$

When the mean value of 0.6 is considered for  $\cos^2 \phi$ , we obtain approximately the same value as in (32). The same mean value is, of course, obtained from (26) when one takes  $\sigma = 0$ ,  $J = 1/R$ , and the mean value of gravity. Equation (26) cannot be used for the evaluation of the normal gravity gradient inside the reference ellipsoid, since the normal mass distribution  $\sigma^N$  within the ellipsoid is not defined (cf. §20.3).

The actual gravity field gradient also varies regionally and locally due to the uneven distribution of masses, i.e., due to the topography and the density variations underneath the earth's surface. These density variations, in turn, are reflected in the varying mean curvature  $J$  of the corresponding equipotential surfaces (cf. FIG. 11). Regionally valid values of  $J$  could be obtained from the regional shape of the geoid. Local values have to be determined directly. Neither the radius of curvature nor its reciprocal  $J$  can be observed; there are, however, techniques available whereby one measures other quantities and  $J$  is then derived from those. One such quantity is the deflection of the vertical.

Other local features of the actual gravity field can be measured directly by *torsion*

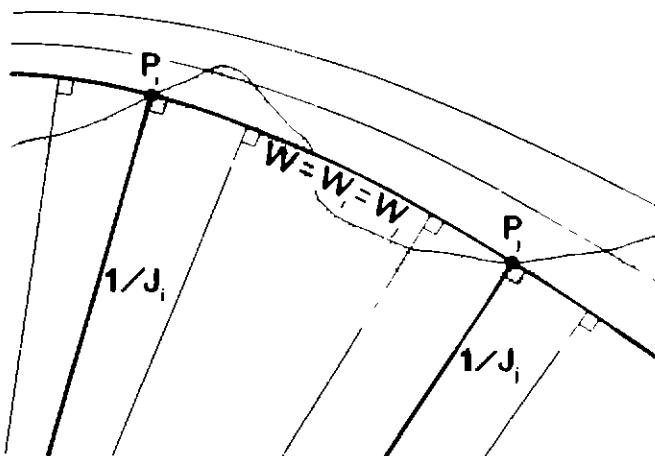


FIG. 21.11. Curvature of equipotential surfaces.

*balance.* Using a technique developed by Eötvös, we can obtain the following quantities:

$$\frac{\partial g}{\partial x}, \quad \frac{\partial g}{\partial y}, \quad \frac{\partial^2 W}{\partial x \partial y}, \quad \frac{\partial^2 W}{\partial y^2} - \frac{\partial^2 W}{\partial x^2} \quad (21.36)$$

in the coordinate system shown in FIG. 10 [MUELLER, 1963].

For the determination of  $\partial g / \partial H$  inside the earth, hypotheses concerning the density distribution have to be relied upon. Many such hypotheses have been proposed and are used for different purposes. In this section, we will show two such gradients that are generally used: the first is attributed to Poincaré and Pray, and the second was developed by Bouguer.

(a) The *Poincaré–Pray gradient* is based on the assumption that the product  $gJ$  for the actual gravity field is, on average, the same as  $\gamma J^N$  for the normal field [HEISKANEN AND MORITZ, 1967]. Under this assumption, subtracting (27) from (26), we arrive at

$$\frac{\partial g}{\partial H} \doteq \frac{\partial \gamma}{\partial H} + 4\pi G\sigma. \quad (21.37)$$

Then taking  $\partial \gamma / \partial H \doteq -0.3086 \text{ mGal/m}$  and an average density  $\sigma = 2.67 \text{ g cm}^{-3}$ , we obtain

$$\frac{\partial g}{\partial H} \doteq -0.0848 \text{ mGal/m.}$$

(21.38)

This gradient is used in evaluating Helmert's orthometric heights (cf. §16.4) and can be considered a good approximation for the layers immediately underneath the surface of the earth, as witnessed by, e.g.,  $-81 \mu\text{Gal/m}$  observed by McCULLOH [1965] or STRANGE [1982].

(b) To comprehend the basic idea of Bouguer's gradient, let us have a look at FIG. 12. To determine the gravity gradient between the terrain point *A* and the corre-

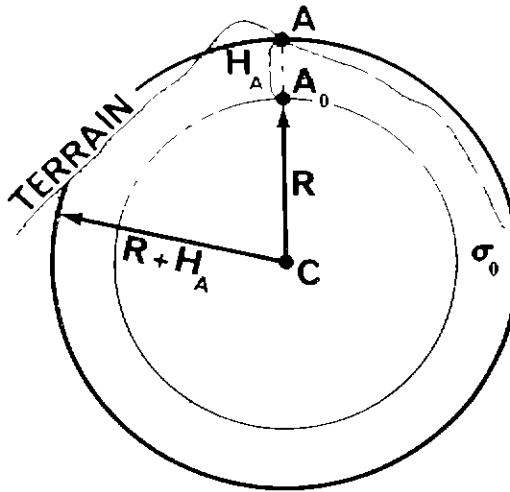


FIG. 21.12. Bouguer's gradient.

sponding point  $A_0$  on the geoid, let us begin by assuming that the geoid is a sphere of radius  $R$ . The gradient is then evaluated in two steps: first, the part of the gradient due to this assumed geoid is obtained, and then the gradient of the spherical shell  $\mathcal{E}$  of thickness  $H_A$  and uniform density  $\sigma_0$  is determined. The first step again leads to the free air gradient, if laterally homogeneous distribution of masses is assumed. The second step is slightly more involved.

The gravitational acceleration  $\bar{g}_g^{\mathcal{E}}$  due to the shell, expressed in geocentric spherical coordinates, satisfies at  $A$  the following equation (cf. (20.7) and (20.14)):

$$\operatorname{div} \bar{g}_g^{\mathcal{E}} = \nabla \cdot \bar{g}_g^{\mathcal{E}}(r, \theta, \lambda)|_A = -2\pi G \sigma_0. \quad (21.39)$$

Since the shell is considered spherical, then  $\bar{g}_g^{\mathcal{E}}$  is, in our coordinate system, a function of  $r$  only, and the derivatives with respect to  $\theta$  and  $\lambda$  disappear. With the realization that the direction of  $\bar{g}_g^{\mathcal{E}}$  runs opposite to the direction of  $r$  (hence the change of sign), (39) reduces to

$$\frac{2}{r_A} g_g^{\mathcal{E}}(r_A) + \left. \frac{\partial g_g^{\mathcal{E}}}{\partial r} \right|_A = 2\pi G \sigma_0, \quad (21.40)$$

where the second term is the gradient we seek.

The gravitational acceleration  $g_g^{\mathcal{E}}$  due to the shell is easily determined, since the field of the shell is radial. We get

$$g_g^{\mathcal{E}}(r_A) = \frac{GM^{\mathcal{E}}}{(R + H_A)^2}, \quad (21.41)$$

where  $M^{\mathcal{E}}$  is given by

$$M^{\mathcal{E}} = \frac{4}{3}\pi\sigma_0 [(R + H_A)^3 - R^3] \doteq 4\pi\sigma_0 R^2 H_A. \quad (21.42)$$

The gradient at  $A$  thus is equal to

$$\frac{\partial g_g^e}{\partial H} \Big|_A \doteq +2\pi G\sigma_0 - 8\pi G\sigma_0 \frac{H_A}{R}. \quad (21.43)$$

The first term is called the *Bouguer plate gradient*. It can be obtained simply as a gradient produced by a straight plate of density  $\sigma_0$  that extends into infinity. Selecting a local Cartesian coordinate system as shown in FIG. 13, for the plate's gravitational acceleration  $\bar{g}_g^p$  at  $A$  we can immediately write

$$\nabla \cdot \bar{g}_g^p = -2\pi G\sigma_0, \quad (21.44)$$

and since the field does not depend on either  $x$  or  $y$ , we get the expression for the first term in (43). Logically, the second term in (43) originates from the fact that the shell is merely the plate wrapped around the geoid. It is thus called the *curvature gradient*. It is curious to note that the curvature gradient is very small (+0.00012 mGal/m for each kilometre of  $H$ ) and, as such, is usually neglected; the spherical shell thus produces on the surface of the shell a gradient that is, for all practical purposes, equivalent to that of Bouguer's plate.

Having determined the gradient of the shell, we must add it to the free air gradient. Assuming the crustal density  $\sigma_0$  to be  $2.67 \text{ g cm}^{-3}$ , we finally obtain the complete *Bouguer gradient* as

$$\frac{\partial g}{\partial H} \doteq -0.1967 \text{ mGal/m.} \quad (21.45)$$

To close this section, let us make a few observations pertaining to the gravity gradients. First, note that the Bouguer gradient is exactly the mean between the free air and Poincaré–Pray gradients. The explanation of this fact becomes clear when we look at all three of these gradients as divergences of  $\bar{g}$ . While the free air gradient is evaluated as the divergence of  $\bar{g}$  in empty space ( $\sigma=0$ ), and the Poincaré–Pray gradient is evaluated as the divergence of  $\bar{g}$  inside the earth, the Bouguer gradient is evaluated on the surface of the earth (for  $\sigma/2$ ). Thus the Bouguer gradient is realistic neither inside nor outside the earth and is seldom used in geodesy. It is, however,

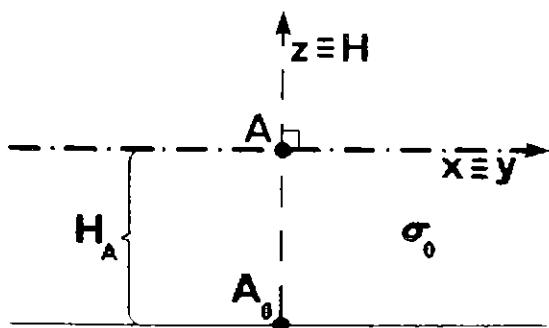


FIG. 21.13. Bouguer's plate.

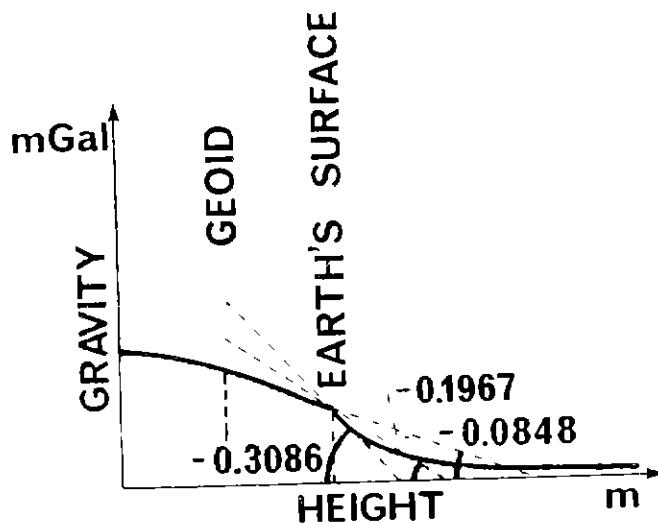


FIG. 21.14. Vertical gradient of gravity.

very useful for various geophysical tasks and, as such, is used there extensively. The relation of the three gradients on, above, and underneath the earth's surface is shown in FIG. 14.

Other models for the vertical gradient of gravity have been suggested by scores of researchers, but none has found wide application. The interested reader is referred to LAMBERT [1930], and HEISKANEN AND MORITZ [1967].

### 21.3. Curvature of the plumb line

Another important gravity field characteristic of a local nature is the curvature of the plumb line, actual or normal, whose role has also been mentioned on several occasions (e.g., §6.4 and §21.1). To investigate how this quantity is related to other field parameters, let us first have a look at the situation as it is depicted in FIG. 15. What we want to study is the angle  $\delta\epsilon$  or rather its projections onto the meridian and prime vertical planes.

The mathematical technique used in this investigation parallels that used in the previous section. First, a convenient local right-handed Cartesian coordinate system is selected so that the  $z$ -axis is tangent to the plumb line at the point of interest  $A$  and positive in an upward direction. Then, the  $x$ -axis is specified to point north so that we again have the equivalent of an LA system with the direction of the  $y$ -axis reversed.

Let us concentrate on the curvature in the meridian plane ( $x, z$ ) first—see FIG. 16. The equation of the projection of the plumb line in the  $(x, z)$  plane can be written as

$$x = x(z). \quad (21.46)$$

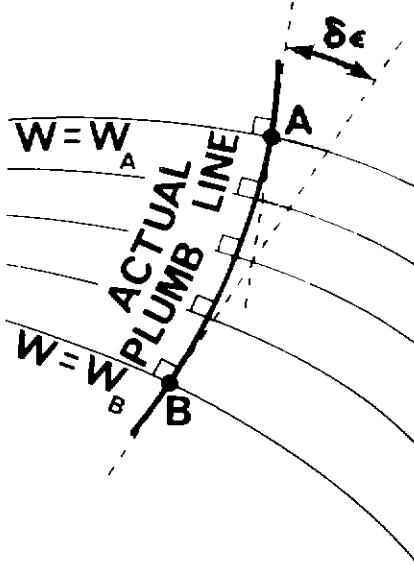


FIG. 21.15. Curvature of the plumb line.

It is quite apparent that this function has an extreme at  $A$ , i.e.,

$$\frac{dx}{dz} \Big|_A = 0. \quad (21.47)$$

Thus the curvature  $k_x$  of the projection of the plumb line is given by (cf. the reasoning following (25))

$$k_x \Big|_A = \frac{d^2x}{dz^2} \Big|_A. \quad (21.48)$$

To evaluate the second derivative at  $A$ , it is expedient to formulate the differential equation of the plumb line first. This can be easily done by taking an infinitesimally short vector  $d\bar{a}$ ,

$$d\bar{a} = dx \bar{i} + dy \bar{j} + dz \bar{k}, \quad (21.49)$$

and making it parallel to the actual gravity vector  $\bar{g}$ :

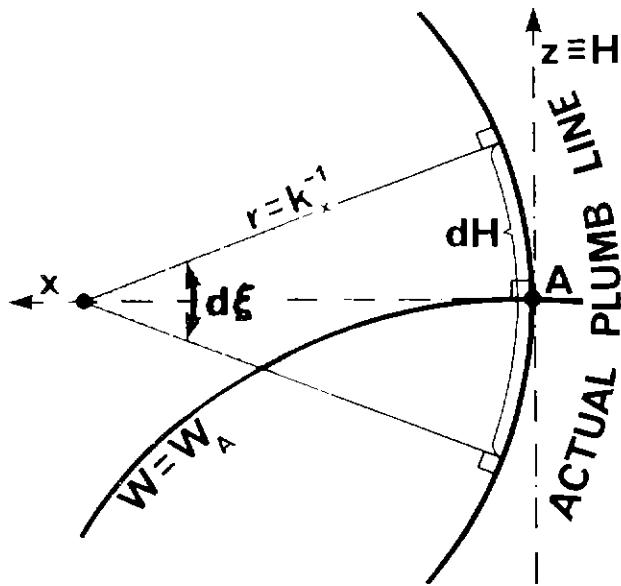
$$\bar{g} = \nabla W = \frac{\partial W}{\partial x} \bar{i} + \frac{\partial W}{\partial y} \bar{j} + \frac{\partial W}{\partial z} \bar{k}. \quad (21.50)$$

We get

$$dx / \frac{\partial W}{\partial x} = dy / \frac{\partial W}{\partial y} = dz / \frac{\partial W}{\partial z},$$

(21.51)

which is the differential equation being sought (cf. §3.3). The first derivative of  $x$

FIG. 21.16. Local Cartesian system (LA-system with  $y$ -axis inverted).

with respect to  $z$  is evidently equal to

$$\frac{dx}{dz} = \frac{\partial W}{\partial x} / \frac{\partial W}{\partial z}. \quad (21.52)$$

The second derivative is then obtained as the total derivative of the first derivative (see §3.2)

$$\frac{d^2x}{dz^2} = \left( \frac{\partial W}{\partial z} \right)^{-2} \left[ \frac{\partial W}{\partial z} \left( \frac{\partial^2 W}{\partial x \partial z} + \frac{\partial^2 W}{\partial x^2} \frac{dx}{dz} \right) - \frac{\partial W}{\partial x} \left( \frac{\partial^2 W}{\partial z^2} + \frac{\partial^2 W}{\partial z \partial x} \frac{dx}{dz} \right) \right]. \quad (21.53)$$

Because the  $x$ -axis is tangent to the equipotential surface  $W = W_A$  (cf. FIG. 16), we get  $\partial W / \partial x = 0$ , and, considering (47), (53) reduces to

$$k_x|_A = \frac{\partial^2 W}{\partial x \partial z} / \frac{\partial W}{\partial z}|_A. \quad (21.54)$$

Assuming  $W$  along the plumb line to be sufficiently smooth, we can interchange the arguments in the second derivative and finally obtain

$$k_x|_A = \frac{1}{g} \frac{\partial g}{\partial x}|_A. \quad (21.55)$$

To convert the curvature into the differential change  $d\xi$  of the meridian deflection component  $\xi$ , the obvious equation (see FIG. 16) is

$$d\xi = -k_x dH, \quad (21.56)$$

where the minus sign again reflects the sign convention for the deflection components (cf. (18)), as the reader can verify. The overall change  $\delta\xi$  in  $\xi$  due to the

curvature of the plumb line between  $A$  and  $B$  (cf. FIG. 15) is then obtained by integrating the differential changes between  $A$  and  $B$ :

$$\delta\xi = - \int_A^B \frac{1}{g} \frac{\partial g}{\partial x} dH. \quad (21.57)$$

A completely analogous formula can be derived for the prime vertical component:

$$\delta\eta = + \int_A^B \frac{1}{g} \frac{\partial g}{\partial y} dH. \quad (21.58)$$

It is illustrative to apply these formulae now to the normal gravity field above the geocentric ellipsoid. It can be seen immediately that  $\partial\gamma/\partial y=0$ , since there are no changes in the normal gravity in the east-west direction; the normal field is symmetrical and does not depend on  $\lambda$ . Hence,

$$\delta\eta^N = 0. \quad (21.59)$$

On the other hand, the normal gravity does change in the north-south direction since the normal field depends on  $\phi$ . Taking just the first two terms in (20.85), we can write the normal gravity in the following form:

$$\gamma(\phi, h) = \gamma_0(\phi) + \frac{\partial\gamma}{\partial h} h \doteq \gamma_E(1 + f \sin^2\phi) + \frac{\partial\gamma}{\partial h} h. \quad (21.60)$$

In the first approximation,  $\partial\gamma/\partial h = \partial\gamma/\partial H$  is constant (cf. (30)), and we obtain

$$\frac{\partial\gamma}{\partial x} = \frac{1}{R(\phi)} \frac{\partial\gamma}{\partial\phi} \doteq \frac{\gamma_E}{R} f \sin 2\phi. \quad (21.61)$$

For the region directly above the reference ellipsoid, we get

$$\delta\xi^N \doteq - \int_A^B \frac{\gamma_E}{\gamma R} f \sin 2\phi dH \doteq 0.17'' \sin 2\phi \Delta H_{AB}, \quad (21.62)$$

where  $\Delta H_{AB}$  is expressed in kilometres.

Note that the curvature of the normal plumb line is fairly small; moreover, it is only weakly related to the curvature of the actual plumb line. Conversely, the actual plumb line curvature is much more irregular and also much more pronounced. Hence, contrary to the normal curvature, the evaluation of the effect of the actual curvature is a very difficult task. To calculate this effect, let us take a point  $A$  on the surface of the earth and another differentially close point  $dS$  apart. Then the projection of the change  $\delta\epsilon$  in the deflection  $\epsilon$  due to the curvature of the projected plumb line into the vertical plane passing through these two points is equal to

$$\delta\epsilon = \frac{dH - dl}{dS}, \quad (21.63)$$

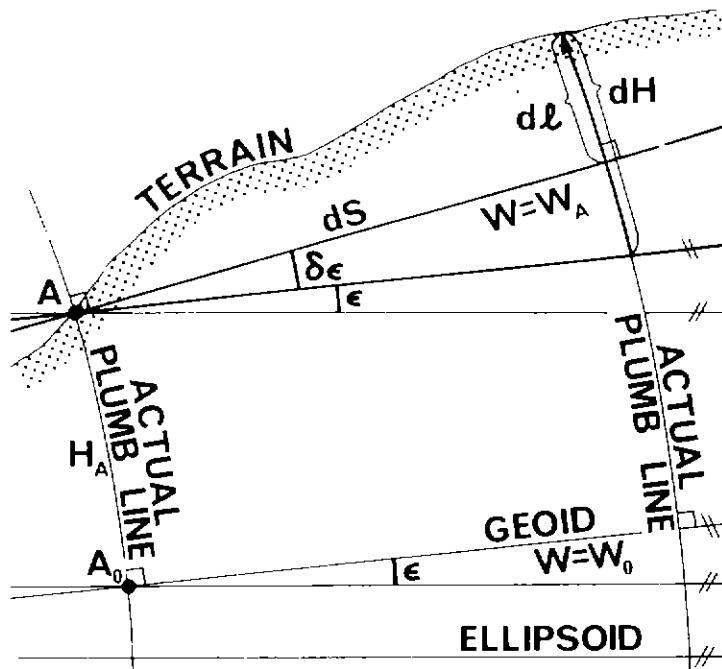


FIG. 21.17. Relation between curvature of plumb line and height.

where  $dH$  is the differential increment in orthometric height along  $dS$ , and  $dl$  is the increment in observed height along  $dS$  (see FIG. 17). But,  $dH - dl$  is simply the change in orthometric correction OC as seen in §16.4 along  $dS$ , i.e.,  $dOC$ .

From the definition of orthometric height, in a few simple steps we can derive

$$dOC = H_A \frac{\bar{g}' - \bar{g}'}{\bar{g}'} + dl \frac{\bar{g} - \bar{g}'}{\bar{g}'} , \quad (21.64)$$

where  $\bar{g}'$  is the mean gravity of  $A$ , and  $\bar{g}'$  is the mean gravity for the second point (both in the sense of the definition of orthometric heights). The mean surface gravity between the two points is denoted by  $\bar{g}$ .

Let us now write  $\bar{g}' = g_A - \frac{1}{2}(\partial g / \partial H)H_A$ ,  $\bar{g} = g_A + \frac{1}{2}(\partial g / \partial S)dS$ , and  $\bar{g}' = g - \frac{1}{2}(\partial g / \partial H)H$ . Further, let us assume  $\partial g / \partial H$  to be constant, or in other words,  $\partial^2 g / (\partial H \partial S)$  to be equal to zero. Then (64) becomes

$$dOC = \frac{\frac{\partial g}{\partial H} H dH - \frac{\partial g}{\partial S} H dS}{2\bar{g}} . \quad (21.65)$$

The deflection of the vertical in the desired direction is then affected by

$$\delta\epsilon = -\frac{H}{2\bar{g}} \left( \frac{\partial g}{\partial S} - \frac{\partial g}{\partial H} \frac{dH}{dS} \right) ,$$

(21.66)

where all the quantities except  $\partial g / \partial H$  can be observed on the surface of the earth. More general formulae that do not require the assumption of constancy of the

vertical gravity gradient can be found in BODEMÜLLER [1957]. Alternative approaches have been suggested by other researchers; we shall see one viable alternative in Chapter 22.

To compare the curvature of the actual plumb line (between the geoid and the terrain) with that of the normal plumb line, let us take the following case:  $\partial g/\partial S = -50 \mu\text{Gal}/\text{m}$  (compare with  $|\partial\gamma/\partial S| \leq 0.5 \mu\text{Gal}/\text{m}$ ),  $\partial g/\partial H = -85 \mu\text{Gal}/\text{m}$ , and  $dH/dS = 0.2$ , where the two terms in (66) tend to cancel each other because of the negative correlation between  $g$  and  $H$ . We obtain  $\delta\epsilon \doteq 3.3''$  per 1 km of elevation, which agrees in the order of magnitude with the amount reported by KOBOLD AND HUNZIKER [1962] for the Alps.

The horizontal gradient of gravity  $\partial g/\partial S$  needed in (66) can be determined either from gravity observations, existing maps, or from direct measurements. The horizontal gradient can be observed by means of gradiometers [FORWARD, 1974] or torsion balance [MUELLER, 1963]. The accuracy of (66) is somewhat questionable because of the uncertainty in  $\partial g/\partial H$ .

## 21.4. Topographical and isostatic effects

The local behaviour of the actual field reflects the local and regional irregular distribution of masses. The most conspicuous irregularity is caused by the irregular shape of the earth's surface and, to a less evident degree, by isostasy. Therefore, these two phenomena deserve special attention in this section.

As was seen in §8.2, the earth's crust is in a state of isostatic equilibrium over most of the earth's surface. This means that the geoid, being an equipotential surface, should not have its shape much affected by the presence of irregular topography; the effect of redundant masses above the geoid is compensated for by a density deficiency deep underground. An inverse situation occurs on the seas; there the deficiency of superficial masses is compensated for by redundancy of dense masses underground. The shape of the equipotential surfaces and plumb lines is shown schematically in FIG. 18.

On the other hand, our observations on the surface of the earth are affected very strongly by topography; the more so the further we are removed from the geoid, i.e., the higher the altitude. One example of such an effect is seen in (66):  $dH/dS$  is simply the slope of the terrain. We can say that the second term in (66) describes the *topographical effect on the curvature of the actual plumb line*. That does not mean, obviously, that the curvature is zero in flat country.

Evidently, the terrain must also exert an effect on gravity. To show this, let us have a look at FIG. 19. It is easy to see that, compared with a horizontal surface, the presence of topography (slope in our case) in the vicinity of  $A$  produces an attraction acceleration that always points upward, no matter whether the topography represents mass redundancy or deficiency. Thus, the presence of topography accounts for a decrease of gravity at  $A$ ; generally, the gravity value observed on the earth's surface is always smaller than it would be if the earth's surface were absolutely

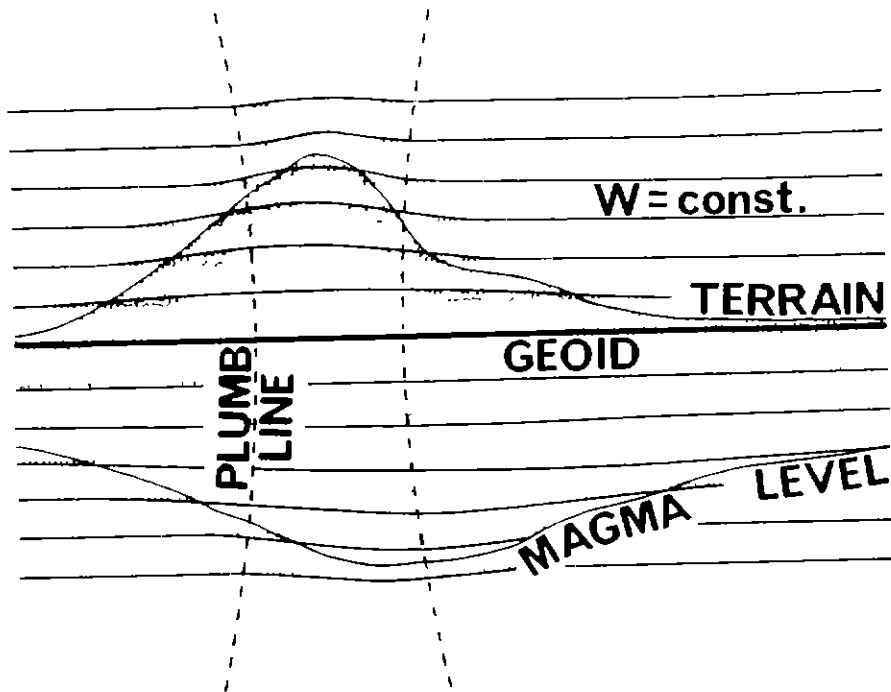


FIG. 21.18. Topographical and isostatic effects.

horizontal. The equipotential surfaces thus become more spread out in the presence of topography causing the ripples seen in FIG. 18.

This phenomenon is most expediently evaluated through the *topographical effect*  $\delta g^T$  on observed gravity. Because topography is generally irregular, its effect cannot be expressed in an analytical form, and numerical integration has to be used over the earth's surface. For this purpose, the surface of the earth is divided into compartments (cells), the contribution of topography within each cell is assessed individu-

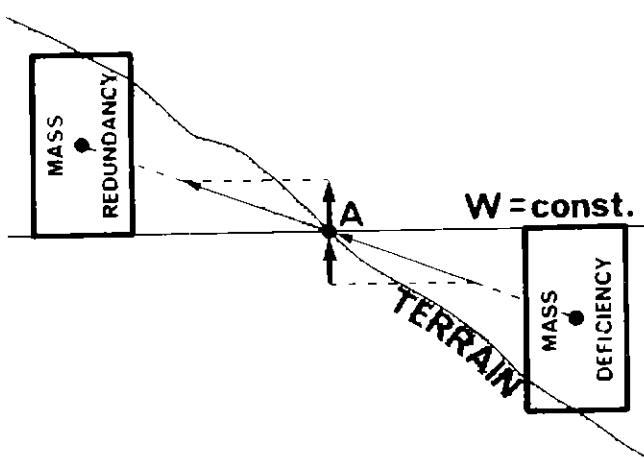


FIG. 21.19. Effect of topography on observed gravity.

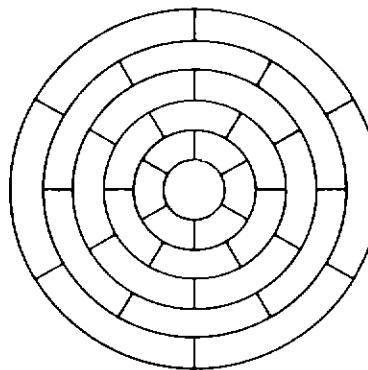


FIG. 21.20. Circular template.

ally, and then the contributions of all the compartments are added together to give the overall effect. Since a similar numerical integration is often used in geodesy in other contexts as well (see Chapter 22), it will serve a useful purpose to show here how the pertinent formulae are derived.

FIG. 20 portrays one possible shape of a template that can be used to divide the surface into surface cells. Templates of different shapes may be used; the circular template, combined with the cylindrical coordinate system, is the most versatile for manual computations. To evaluate the contribution of one cell of such a circular template, let us turn to FIG. 21 that depicts one such typical compartment of average height  $\Delta H$  above the observation point  $A$ . Its overall attraction acceleration  $\delta f$  can be obtained by integrating the infinitesimal attraction accelerations  $d\delta f$  exerted by all mass elements  $dm$  within the compartment. Then the gravity effect  $\delta g^T$  of the compartment is derived by projecting  $\delta f$  onto the (vertical)  $z$ -axis.

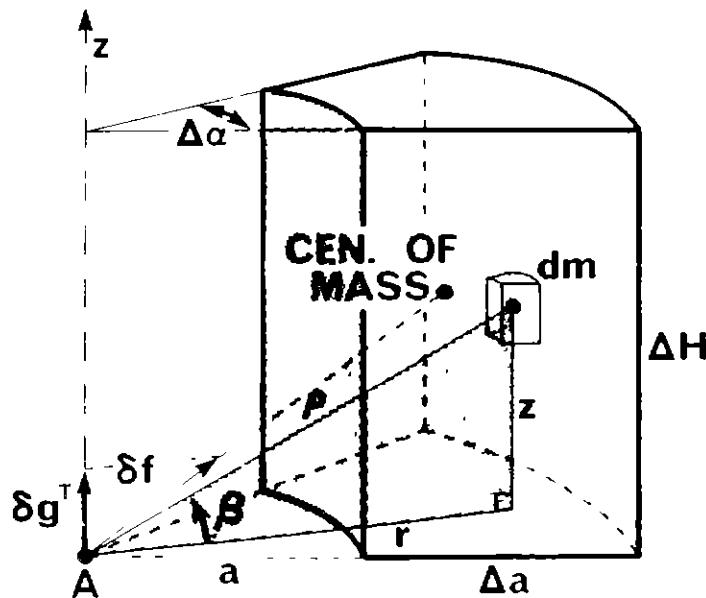


FIG. 21.21. Attraction of one compartment.

In our derivation, we will do it a little differently; we will evaluate the differential gravity effects  $d\delta g^T$  first, and then integrate them. The differential attraction acceleration is evidently given by

$$d\delta f = G \frac{dm}{\rho^2}, \quad (21.67)$$

where  $\rho^2 = r^2 + z^2$ . The corresponding differential gravity effect is

$$d\delta g^T = d\delta f \sin \beta = G \frac{\sigma dV}{(r^2 + z^2)} \frac{z}{(r^2 + z^2)^{1/2}}. \quad (21.68)$$

Now, considering the density  $\sigma$  of the whole compartment  $\Delta a \times \Delta H \times a \Delta \alpha$  to be constant, we can write the total gravity effect:

$$\delta g^T = G \sigma \int_{\alpha=0}^{\Delta \alpha} \int_{z=0}^{\Delta H} \int_{r=a}^{a+\Delta a} \frac{zr}{(r^2 + z^2)^{3/2}} dr dz d\alpha. \quad (21.69)$$

Integration with respect to  $\alpha$  yields

$$\delta g^T = G \sigma \Delta \alpha \int_{z=0}^{\Delta H} \int_{r=a}^{a+\Delta a} \frac{zr}{(r^2 + z^2)^{3/2}} dr dz. \quad (21.70)$$

Using substitution  $\rho^2 = r^2 + z^2$  we get, after integrating with respect to  $\rho$ ,

$$\delta g^T = G \sigma \Delta \alpha \int_a^{a+\Delta a} \left( 1 - \frac{r}{(r^2 + \Delta H^2)^{1/2}} \right) dr. \quad (21.71)$$

This integral can be finally evaluated using the same substitution as above, i.e.,  $\rho^2 = r^2 + \Delta H^2$ , to obtain

$$\delta g^T = G \sigma \Delta \alpha \left[ \Delta a - \sqrt{(a + \Delta a)^2 + \Delta H^2} + \sqrt{a^2 + \Delta H^2} \right]. \quad (21.72)$$

The numerical integration over all the compartments can now be carried out. Denoting the effect of each compartment by two subscripts  $i$  and  $j$ , the first giving the radial position of the appropriate ring, and the second giving the position of the compartment within the ring, we get, assuming uniform density  $\sigma$  throughout,

$$\delta g^T = G \sigma \sum_i \sum_j \Delta \alpha_j \left[ a_{i+1} - a_i + \sqrt{a_i^2 + \Delta H_{j,i}^2} - \sqrt{a_{i+1}^2 + \Delta H_{j,i}^2} \right]. \quad (21.73)$$

Note that the terms in square brackets tend to cancel out as  $a$  increases. Thus we do not have to carry the integration ordinarily beyond a few tens of kilometres. In fact, the above formula is valid only when we do not have to consider the curvature of the earth.

In addition note that if  $\Delta H$  is equal to zero everywhere, i.e., if the surface of the earth is flat, the effect  $\delta g^T$  is also equal to zero, as one would expect. In reality, values reaching over a hundred milligal are experienced in mountains, e.g., HEISKANEN AND VENING MEINESZ [1958]. The topographical effect on observed gravity can thus be very important.

It is revealing to establish the limits of the topographical effect and to have a look at the two extreme cases. Rewriting eqn. (26) for the surface of the earth (cf. eqn. (20.14))

$$\frac{\partial g}{\partial H} = k\pi G\sigma - 2\omega^2 - 2gJ, \quad (21.74)$$

we see immediately that for  $k = 0$  we obtain the free air gradient (valid outside the earth), while for  $k = 4$  the Poincaré–Pray gradient (valid inside the earth) results. Here, the value of  $k$  depends on the shape of the earth's surface (cf. FIG. 20.2). The different situations that may occur are depicted in FIG. 22.

Realizing now that the Bouguer gradient ( $\partial g/\partial H$ )<sup>B</sup> is given by eqn. (74) with

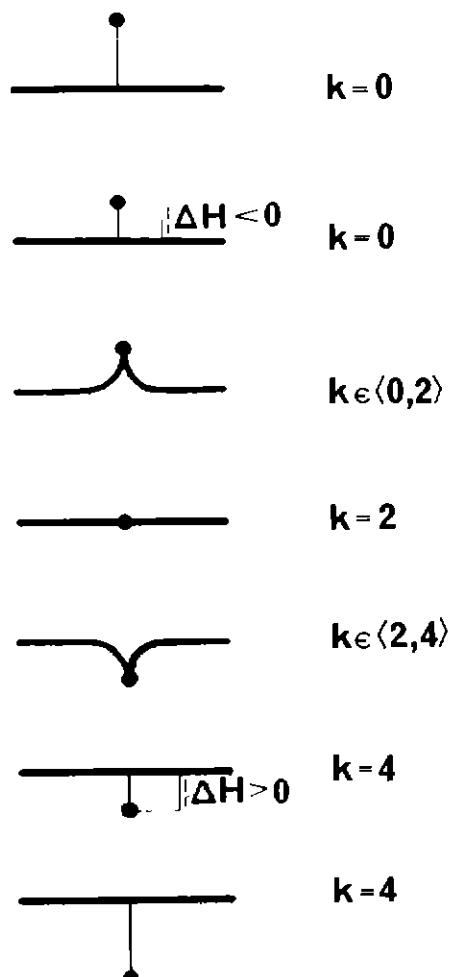


FIG. 21.22. Topographical effect on the vertical gradient of gravity.

$k = 2$  (cf. §21.2), we can rewrite eqn. (74) as

$$\frac{\partial g}{\partial H} = \left( \frac{\partial g}{\partial H} \right)^B + (k - 2)\pi G \sigma, \quad (21.75)$$

where the second term on the right-hand side is the *topographical gravity gradient*  $(\partial g / \partial H)^T$ . It has to be added to the Bouguer gradient to obtain the *total surface gradient*. Clearly, the topographical gradient satisfies the following inequalities:

$$-0.1119 \text{ mGal/m} \leq \left( \frac{\partial g}{\partial H} \right)^T \leq 0.1119 \text{ mGal/m}, \quad (21.76)$$

so that the total surface gradient is indeed between the Poincaré–Pray value of  $-0.0848 \text{ mGal/m}$  and the free air value of  $-0.3086 \text{ mGal/m}$  (see FIG. 23). The free air value is obtained for the ‘needle-like topography’ ( $k = 0$ ), while the Poincaré–Pray value is obtained for the ‘well-like topography’ ( $k = 4$ ).

To evaluate the topographical effect  $\delta g^T$  on observed gravity, the topographical gradient has to be multiplied by  $\Delta H$ , the height of the terrain feature taken relative to the observing point (see FIG. 22). The needle-like topography is characterized by a negative  $\Delta H$ , while the well-like topography gives a positive  $\Delta H$ . In both cases, the product,

$$\delta g^T = \left( \frac{\partial g}{\partial H} \right)^T \Delta H, \quad (21.77)$$

is a positive number, as already shown above.

For geophysical applications, the topographical effect is usually added to the *Bouguer gravity anomaly* which is obtained as

$$\Delta g^{(B)} = g - \left( \frac{\partial g}{\partial H} \right)^B H - \gamma_0, \quad (21.78)$$

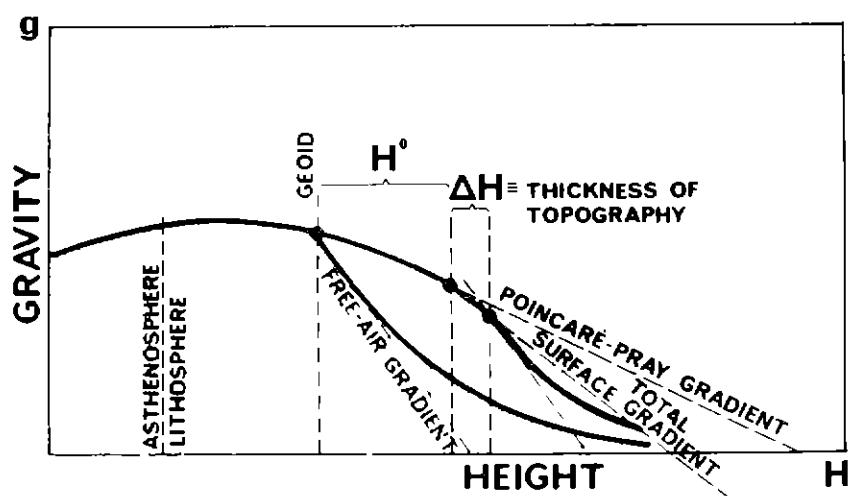


FIG. 21.23. Variations of gravity with height and with topography. (The case of the earth without a topography is shaded.)

where  $g$  is the gravity value observed at point  $P$  on the earth's surface,  $\gamma_0$  is the normal gravity on the ellipsoid (evaluated for the latitude of  $P$ ), and  $H$  is the orthometric height  $H^0$  of  $P$ . The physical meaning of the Bouguer anomaly is obscure. Bouguer anomalies show a considerable broad regional negative correlation with topography [HEISKANEN AND VENING MEINESZ, 1958] indicating that the gravity reduced to the geoid by the Bouguer gradient is too small underneath the mountain. This, in turn, indicates that the Bouguer gradient is too small in absolute value. This will be explained in §22.1.

As was stated at the beginning of this section, the shape of the geoid should not be much influenced by topography. Hence, the correlation of the geoidal gravity anomalies with the topography runs contrary to our expectation, when the situation is viewed from the isostasy point of view: such anomalies indicate an unbalanced distribution of masses which, in turn, violates the physical principle of isostasy. This is because the findings of this section are valid only on and immediately below the surface of the earth. When the behaviour of the gravity field at the geoid level is studied, compensation must be made for the *effect of isostasy* which has not yet been taken into account.

The effect of isostasy is to decrease the value of gravity on the geoid underneath the mountains and increase it on the seas as compared with what the values would be in the absence of isostasy. This means that isostasy tends to decrease the absolute value of the vertical gravity gradient from the free air value toward the Poincaré–Pray value. The effect can be evaluated in a fashion similar to that used for the topographical effect. Provided the density distribution of the crust is known or postulated, the crust can again be divided into compartments. The effect of each compartment is then evaluated separately, and the overall effect is obtained through numerical integration over an appropriate area.

When evaluating the *isostatically compensated gravity anomaly*, it is customary to again use the Bouguer anomalies as the basis. This is because the Bouguer anomalies are readily available in various convenient forms. As to the mass distribution within each compartment, one generally has to use a hypothetical one. All the three models of density distribution seen in §8.2 had been used in practice in many attempts to determine the isostatic effect. HAYFORD AND BOWIE [1912] used Pratt's model with a uniform depth of the crust of 113.7 km and 122 km, a little too thick as we know today. Other values were used by other researchers. Airy's model was used by HEISKANEN [1938] in his investigations; Heiskanen postulated a progression of normal crustal depths of 20, 30, 40, and 60 km. VENING MEINESZ [1939] used depths of 10, 20, 40, 60, and 80 km when working with his own model. Results of investigations of the isostatic effect are usually presented in the form of tables or maps. A comprehensive summary can be found in HEISKANEN AND VENING MEINESZ [1958].

Isostasy, of course, affects all the other parameters of the gravity field as well. We shall discuss the effect on  $N$  and  $\xi$  in Chapter 22, together with the topographical effect, but first it is necessary to develop the proper mathematical tools to handle them. Let us just state here that the allowance for isostasy is much more critical in mountainous areas and over the deep sea. In flat and not too elevated regions, the

isostatic models do not deviate much from stratified models of the earth. For such regions, all the formulae developed in this chapter can be used without worrying about isostasy.

## CHAPTER 22

# DETERMINATION OF THE GRAVITY FIELD FROM GRAVITY OBSERVATIONS

In the preceding two chapters, various theoretical relations that exist among the parameters of the earth's gravity field were discussed, but little was said about ways in which to determine the values of these parameters. Also, no attempt has been made to build the mathematical models linking these parameters to observable quantities. In this chapter, the classical problem of the earth's gravity field—how to obtain full information about the field from observations of gravity magnitude alone—will be discussed.

The first section introduces the standard technique proposed by the English mathematician Stokes (cf. §1.3) in the middle of the nineteenth century. The second section is devoted to a modern approach advocated by the Russian physicist Molodenskij about a hundred years later (cf. §1.4). It is interesting to note that even this approach was probably given its impetus by another Englishman, the geophysicist Jeffreys, in the early 1930s. The third section deals with the ways gravity data are acquired and converted into mean anomalies suitable for evaluating the other parameters being sought. The last section describes how to circumvent some of the problems encountered in the evaluation of the necessary surface integrals.

### 22.1. Stokes's concept

Stokes's classical approach [STOKES, 1849] to the determination of gravity field parameters from gravity magnitude observations is based on the solution of the external boundary value problem for the disturbing potential  $T$ . To show how this approach works, let us begin by solving the hypothetical problem already posed in Chapter 20. If we assume (admitting that it is so far incorrect) that there are no masses outside the geoid, then (20.88) is satisfied everywhere outside the geoid.

Since the value of the disturbing potential  $T$  on the ellipsoid is neither known nor observable, different boundary values have to be used, i.e., a boundary value problem of a different type has to be formulated. This is where the fundamental gravimetric equation (21.14) is used. Using the gradient of normal gravity given by

(21.30) and neglecting the terms of the order of flattening, we can rewrite (21.14) as

$$\Delta g \doteq -\frac{2}{R} T - \frac{\partial T}{\partial H}, \quad (22.1)$$

with  $R$  again denoting the mean radius of the earth. This equation furnishes the boundary values and can be used, in conjunction with (20.88), as a boundary value problem of a mixed type (cf. §3.2) involving both the function  $T$  being sought and its derivative  $\partial T / \partial H$  both referring to the geoid. Clearly, the boundary values  $\Delta g$  can be derived from the observed values of  $g$  and  $H$  through a few simple steps and can be treated as observables. In the literature, this problem is often referred to as the *geodetic boundary value problem*. Its solution can now be attempted.

The solution to the geodetic boundary value problem is most expediently sought in the form of ellipsoidal harmonic series (cf. (20.62))

$$T(u_A, \Theta_A, \lambda_A) = \sum_{n=0}^{\infty} \sum_{m=0}^n q_{nm}(u_A) [A_{nm}(T) \cos m\lambda_A + B_{nm}(T) \sin m\lambda_A] P_{nm}(\cos \Theta_A). \quad (22.2)$$

Developing the boundary value  $\Delta g$  regarded, for the moment, to be on the ellipsoid into an ellipsoidal harmonic series, one obtains

$$\Delta g(\Theta_A, \lambda_A) = \sum_{n=0}^{\infty} \sum_{m=0}^n [A_{nm}(\Delta g) \cos m\lambda_A + B_{nm}(\Delta g) \sin m\lambda_A] P_{nm}(\cos \Theta_A), \quad (22.3)$$

where the coefficients are given by (cf. (20.38) and (20.47))

$$\begin{Bmatrix} A_{nm}(\Delta g) \\ B_{nm}(\Delta g) \end{Bmatrix} = \frac{(n-m)!}{(n+m)!} \frac{2n+1}{2\pi} \iint_{\mathcal{E}} \Delta g(\Theta, \lambda) \begin{Bmatrix} \cos m\lambda \\ \sin m\lambda \end{Bmatrix} P_{nm}(\cos \Theta) d\nu. \quad (22.4)$$

Note that for  $m=0$ ,  $2\pi$  is replaced by  $4\pi$ ; the integration is carried out over the ellipsoid  $\mathcal{E}$ , and  $d\nu$  is again the solid angle element. On this ellipsoid, the radial functions  $q_{nm}(u)$  are all equal to 1; with the exception of the values of the coefficients, the series (2) taken on the ellipsoid becomes identical with (3).

To establish the relation between the two sets of coefficients (so that the solution to the geodetic boundary value problem can be expressed in terms of the boundary value  $\Delta g$ ), eqns. (2) and (3) are substituted into (1). This yields

$$\begin{Bmatrix} A_{nm}(\Delta g) \\ B_{nm}(\Delta g) \end{Bmatrix} = -\frac{2}{R} \begin{Bmatrix} A_{nm}(T) \\ B_{nm}(T) \end{Bmatrix} - \frac{\partial q_{nm}(u)}{\partial H} \begin{Bmatrix} A_{nm}(T) \\ B_{nm}(T) \end{Bmatrix},$$

$$n=0, \dots, \infty; m=0, \dots, n. \quad (22.5)$$

Again, to the accuracy of flattening, the derivative of the ellipsoidal radial term is

equal to that of the spherical radial term, i.e.,

$$\frac{\partial q_{nm}(u)}{\partial H} \doteq \frac{\partial}{\partial r} \left( \frac{a}{r} \right)^{n+1} = - \left( \frac{a}{r} \right)^{n+1} \frac{(n+1)}{r}, \quad (22.6)$$

and on the ellipsoid or the geoid, we get in particular:

$$\frac{\partial q_{nm}(u)}{\partial H} \doteq - \frac{n+1}{R}. \quad (22.7)$$

Hence the desired relation is approximately

$$\begin{Bmatrix} A_{nm}(T) \\ B_{nm}(T) \end{Bmatrix} \doteq \frac{R}{n-1} \begin{Bmatrix} A_{nm}(\Delta g) \\ B_{nm}(\Delta g) \end{Bmatrix}. \quad (22.8)$$

Clearly, for  $n=1$ , the relation is not defined. If a geocentric, properly oriented reference ellipsoid is considered, then the coefficients for  $n=1$ , i.e.,  $A_1(T)$ ,  $A_{1,1}(T)$ ,  $B_1(T)$ ,  $B_{1,1}(T)$ , are all equal to zero anyway, as the interested reader can reason out by following the argument in §20.2. Hence, for this geocentric ellipsoid, we can write

$$T(\Theta_A, \lambda_A) = -RA_0(\Delta g) + \sum_{n=2}^{\infty} \frac{R}{n-1} \sum_{m=0}^n [A_{nm}(\Delta g) \cos m\lambda_A + B_{nm}(\Delta g) \sin m\lambda_A] P_{nm}(\cos \Theta_A), \quad (22.9)$$

where, to the accuracy of flattening,  $\cos \Theta_A$  may be replaced by  $\sin \phi_A$ . The 0th order term,

$$T_0 = -RA_0(\Delta g) = -\frac{R}{4\pi} \oint_{\mathbb{S}} \Delta g d\nu = -R\Delta g_0, \quad (22.10)$$

is simply our old friend from §20.4—a constant error in  $T$  due to an improper assessment of the total mass  $M^N$  of the reference ellipsoid needed in the normal potential. The quantity  $\Delta g_0$  denotes the global mean of the gravity anomalies. It is noteworthy that this global mean can be used to assess the goodness of the value of the mass of the earth. Combining (10) and (20.93), one gets

$$\Delta g_0 = \frac{G\delta M}{R^2}. \quad (22.11)$$

Assuming that  $\delta M=0$  and using the notation of (20.90), we can finally write

$$T(\phi_A, \lambda_A) \doteq \sum_{n=2}^{\infty} \frac{R}{n-1} \Delta g_n(\phi_A, \lambda_A).$$

(22.12)

As the solution must not contain the first-order harmonics, these are sometimes called *forbidden harmonics*. This is another way of saying that the solution exists only for a properly oriented geocentric ellipsoid.

This series solution for the disturbing potential (valid for a geocentric reference ellipsoid with a mass equal to that of the earth) can now be converted to a closed

form. To do so, let us interchange the surface integration (in the expressions for coefficients  $A_{nm}(\Delta g)$ ,  $B_{nm}(\Delta g)$ ) with the summation to obtain

$$\begin{aligned} T(\phi_A, \lambda_A) &\doteq \frac{R}{4\pi} \iint_{\mathcal{E}} \Delta g(\phi, \lambda) \sum_{n=2}^{\infty} \frac{k(2n+1)}{n-1} \\ &\quad \times \sum_{m=0}^n \frac{(n-m)!}{(n+m)!} (\cos m\lambda_A \cos m\lambda + \sin m\lambda_A \sin m\lambda) \\ &\quad \times P_{nm}(\sin \phi_A) P_{nm}(\sin \phi) d\nu, \quad k = \begin{cases} 1 & \text{for } m=0 \\ 2 & \text{for } m>0, \end{cases} \end{aligned} \quad (22.13)$$

where the integration is carried over the unsubscripted arguments  $\phi, \lambda$  on the ellipsoid  $\mathcal{E}$ . Making use of (20.51), we get

$$T(\phi_A, \lambda_A) \doteq \frac{R}{4\pi} \iint_{\mathcal{E}} \Delta g(\phi, \lambda) \sum_{n=2}^{\infty} \frac{2n+1}{n-1} P_n(\cos \psi) d\nu, \quad (22.14)$$

where  $\psi$  is the angular distance between  $(\phi_A, \lambda_A)$  and  $(\phi, \lambda)$ .

In (14), the series is a function of  $\psi$  only. A closed form of the same reads [HOBSON, 1931]

$$\begin{aligned} \sum_{n=2}^{\infty} \frac{2n+1}{n-1} P_n(\cos \psi) &= S(\psi) = 1 + \frac{1}{\sin \frac{1}{2}\psi} - 6 \sin \frac{1}{2}\psi \\ &\quad - 5 \cos \psi - 3 \cos \psi \ln(\sin \frac{1}{2}\psi + \sin^2 \frac{1}{2}\psi). \end{aligned} \quad (22.15)$$

This function is known as *Stokes's function*, and its shape is shown in FIG. 1. Substituting this function back into (14), we finally get the closed solution to the hypothetical boundary value problem in the following form:

$$T(\phi_A, \lambda_A) \doteq \frac{R}{4\pi} \iint_{\mathcal{E}} \Delta g(\phi, \lambda) S(\psi(\phi_A, \lambda_A, \phi, \lambda)) d\nu, \quad (22.16)$$

called *Stokes's integral*. From the point of view of the boundary value problem solution, Stokes's function is simply a Green's function [MYINT-U, 1973]. It may

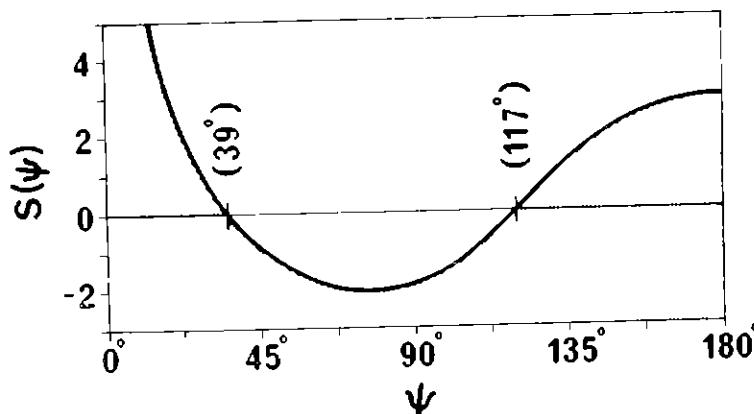


FIG. 22.1. Stokes's function.

also be regarded as an (homogeneous and isotropic) integration kernel: Stokes's integral is thus Green's type of solution (cf. §3.2) to the geodetic boundary value problem.

Even though  $\Delta g$  is a kernel (function of two points: one on the ellipsoid and one on the geoid), the boundary value it represents must be regarded as being formulated on the geoid (and not on the ellipsoid as we stated in the first edition of this book) for the gravity anomaly to represent a proper mixed boundary value in terms of  $T$  [HEISKANEN AND MORITZ, 1967; BOMFORD, 1971]. Just to reiterate, the accuracy of Stokes's integral is of the order of  $f$  (or  $e^2$ ), i.e., 0.3%, due to the various approximations employed. More accurate formulae have been derived by various researchers, e.g., MATHER [1973]. A more serious problem concerns the basic assumption behind setting up the boundary value problem: namely, that there are no masses outside the geoid. Evidently, the boundary value problem as formulated here is improperly posed and can be solved only approximately at the cost of introducing further assumptions. Before worrying about the ways and means to mathematically account for the redundant masses above the geoid, let us look at the other field parameters we seek: namely, the geoidal height and the deflection of the vertical.

If we assume that the problem of redundant masses is eliminated, e.g., by some appropriate correction to  $\Delta g$ , then the disturbing potential given by (16) can be transformed to the geoidal height simply through Bruns's formula (21.4):

$$N(\phi_A, \lambda_A) \doteq \frac{R}{4\pi\gamma_0} \iint_{\mathcal{E}} \Delta g(\phi, \lambda) S(\psi) d\nu, \quad (22.17)$$

where  $\gamma_0$  can be replaced by the mean gravity  $\bar{g}$  (cf. §6.1) without any further deterioration in accuracy. This equation is known as *Stokes's formula*, and it is the most important formula in this chapter. An investigation by the reader will show that if the geocentric reference ellipsoid has a mass different from that of the earth by  $\delta M$ , then the geoidal heights become

$$\begin{aligned} N + N_0 &= -\frac{G\delta M}{R\gamma_0} + \frac{R}{4\pi\gamma_0} \iint_{\mathcal{E}} \Delta g S(\psi) d\nu \\ &= -\frac{R\Delta g_0}{\gamma_0} + \frac{R}{4\pi\gamma_0} \iint_{\mathcal{E}} \Delta g S(\psi) d\nu. \end{aligned} \quad (22.18)$$

Note that each milligal by which the global mean of the gravity anomalies departs from zero causes a constant error of about 6.4 m in geoidal height.

The geoidal deflection of the vertical can be obtained from the geoidal height by substitution for  $N$  in (21.18) to get

$$\begin{aligned} \xi(\phi_A, \lambda_A) &= -\frac{1}{4\pi\gamma_0} \iint_{\mathcal{E}} \Delta g \left. \frac{\partial S(\psi)}{\partial \phi} \right|_A d\nu, \\ \eta(\phi_A, \lambda_A) &= -\frac{1}{4\pi\gamma_0 \cos \phi_A} \iint_{\mathcal{E}} \Delta g \left. \frac{\partial S(\psi)}{\partial \lambda} \right|_A d\nu. \end{aligned} \quad (22.19)$$

The derivatives here can be evaluated through the chain rule where the implicit dependence of  $\psi$  on  $\phi$  and  $\lambda$  is given by (20.50) after  $\theta$  is replaced by  $\frac{1}{2}\pi - \phi$ . We get

$$\begin{aligned} \left. \frac{\partial S(\psi)}{\partial \phi} \right|_A &= \frac{\cos \phi_A \sin \phi - \sin \phi_A \cos \phi \cos(\lambda - \lambda_A)}{-\sin \psi} \frac{dS}{d\psi}, \\ \left. \frac{\partial S(\psi)}{\partial \lambda} \right|_A &= \frac{-\cos \phi_A \cos \phi \sin(\lambda - \lambda_A)}{-\sin \psi} \frac{dS}{d\psi}. \end{aligned} \quad (22.20)$$

On the other hand, rotating  $\bar{u}_p$  to  $\bar{u}$  (see FIG. 2) gives

$$\bar{u} = \mathbf{R}_2(\phi_A) \mathbf{R}_1(-\alpha) \mathbf{R}_2\left(\frac{\pi}{2} - \psi\right) \bar{u}_p. \quad (22.21)$$

The second component yields

$$\sin \psi \sin \alpha = -\cos \phi \sin(\lambda - \lambda_A). \quad (22.22)$$

Multiplying the first component by  $\sin \phi_A$ , the third by  $-\cos \phi_A$ , and adding them together gives us

$$\sin \psi \cos \alpha = \cos \phi_A \sin \phi - \sin \phi_A \cos \phi \cos(\lambda - \lambda_A). \quad (22.23)$$

Putting all these equations together, one finally obtains

$$\begin{Bmatrix} \xi(\phi_A, \lambda_A) \\ \eta(\phi_A, \lambda_A) \end{Bmatrix} = \frac{1}{4\pi\gamma_0} \oint \Delta g(\phi, \lambda) \begin{Bmatrix} \cos \alpha \\ \sin \alpha \end{Bmatrix} \frac{dS(\psi)}{d\psi} d\nu. \end{Bmatrix} \quad (22.24)$$

These equations are usually referred to as the *Vening Meinesz formulae*. Again, the normal gravity may be replaced by the mean gravity  $\bar{g}$  without any detrimental effect

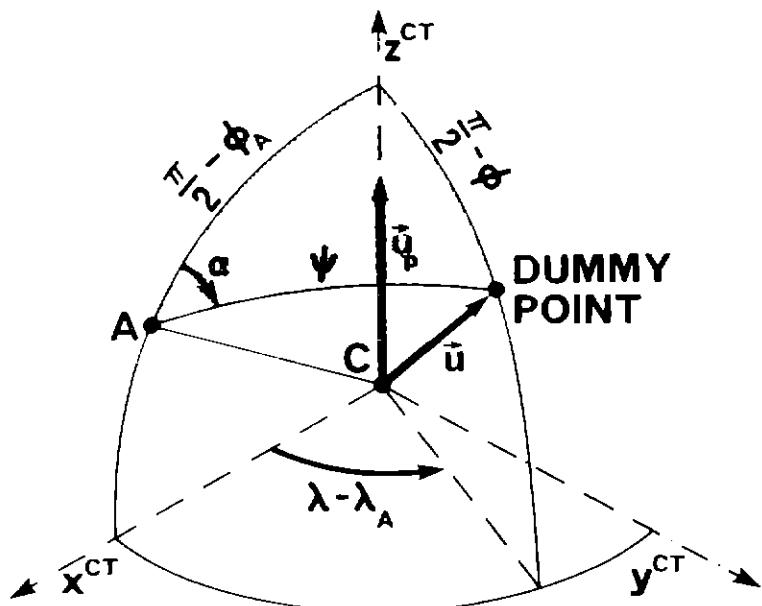


FIG. 22.2. Spherical triangle.

on accuracy. The derivative of Stokes's function is a (anisotropic) kernel, known as the *Vening Meinesz function*, and it can be written in a closed form as [VENING MEINESZ, 1928]

$$\begin{aligned} \frac{dS(\psi)}{d\psi} = & -\frac{\cos \psi}{2 \sin^2 \frac{1}{2}\psi} + 8 \sin \psi - 6 \cos \frac{1}{2}\psi \\ & - 3 \frac{1 - \sin \frac{1}{2}\psi}{\sin \psi} + 3 \sin \psi \ln(\sin \frac{1}{2}\psi + \sin^2 \frac{1}{2}\psi). \end{aligned} \quad (22.25)$$

Its graph is shown in FIG. 3. It should be evident that the constant error  $T_0$  has no effect on the deflection components.

At last we can turn to the question of the validity of the above results in view of the neglected redundant masses. The usual way to overcome the problem arising from the redundant masses is to compensate for their effect by appropriately altering the gravity anomaly  $\Delta g$ . How does  $\Delta g$  react to the removal (i.e., neglect) of the undesired masses? One discovers that neglecting these masses results in the reduction of surface gravity  $g$  and thus also the reduction of  $\Delta g$ . At point  $A$ , this reduction is equal, in the first approximation, to the attraction of Bouguer's plate (cf. eqn. (21.44)) of thickness  $H_A^0$  and density  $\sigma$ :

$$\delta g^P(\phi_A, \lambda_A) = 2\pi G\sigma H_A^0. \quad (22.26)$$

We note that the complete expression also should include the topographical effect  $\delta g^T$  (cf. eqn. (21.73)) and the curvature effect (cf. eqn. (21.43)) both of which are generally fairly small.

To compensate for the neglected masses, let us replace them by an infinitely thin layer of density  $\sigma^S = H^0\sigma$  (cf. §20.4) on the geoid. This layer does not violate the no-masses-outside-the-geoid requirement and can be regarded as the result of 'squashing' the undesired masses onto the geoid. Curiously, the layer can be shown to have almost the same effect on  $\Delta g$  as the neglected masses have had. From FIG. 7

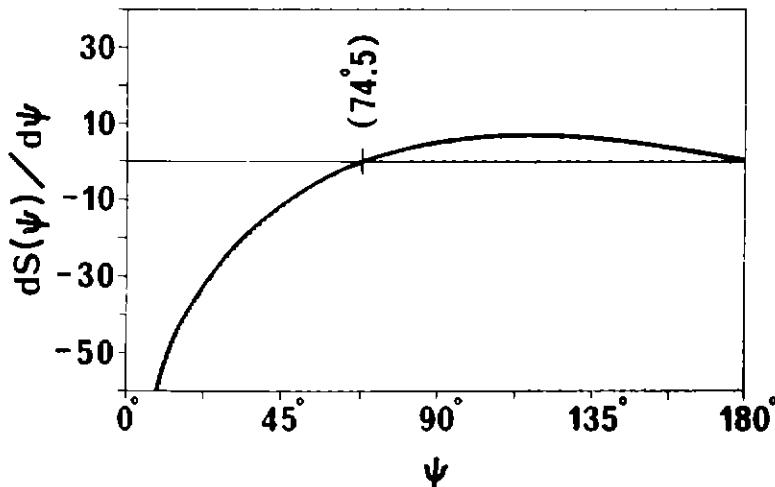


FIG. 22.3. Vening Meinesz's function.

we can express the attractive effect at  $A$  of a differential patch  $d\bar{S}$  of the layer as:

$$dg^L = \frac{G\sigma H^0}{\rho^2} \cdot \frac{H_A^0}{\rho} d\bar{S}. \quad (22.27)$$

Neglecting again as above, the topographical effect (i.e., putting  $H^0 = H_A^0$ ) and the curvature effect (i.e., considering the geoid to be an infinite plane), the attraction of the layer becomes:

$$\delta g^L(\phi_A, \lambda_A) \doteq \int_{\alpha=0}^{2\pi} \int_{r=0}^{\infty} \frac{G\sigma(H_A^0)^2 r}{\rho^3} dr d\alpha, \quad (22.28)$$

where  $d\bar{S}$  has been expressed as  $r dr d\alpha$ . Integration first with respect to  $\alpha$  and then with respect to  $r$  yields an expression identical to eqn. (26): at  $A$  the effect of the infinite plane of density  $\sigma H_A^0$  and the Bouguer plate of thickness  $H_A^0$  and density  $\sigma$  are the same. Rigorously, when replacing masses outside the geoid by the appropriate layer, we should account for the differences in the topographical effects and curvature effects. These differences, however, are even smaller than the effects themselves.

Replacement of the undesired masses by the layer amounts to regarding the gravity station located on the surface of the earth as being suspended in the air. In other words, this situation is equivalent to the case of negative topography of uniform thickness discussed in §21.4. There it was discovered that in such a case, the vertical gradient of gravity becomes equal to the free air gradient, and thus the appropriate anomaly here is simply our old friend from §6.2—the free air anomaly. The effect of the layer on the free air gradient is negligible.

The question now is: What happens to the potential of the earth, and thus to other parameters as well, when the redundant masses are simply disregarded? Not much! As explained in §21.4, at the level of the geoid the effect of those redundant masses on the shape of the geoid is fairly well compensated for by the effect of the irregular mass distribution underneath the geoid caused by the isostasy. In deriving the formula for the free air gradient, a stratified, laterally homogeneous distribution of the masses was assumed, and thus isostasy was neglected; these two neglects balance out. Just how this balance works is shown in FIG. 4. We can see that a reduction of the gravity observed on the earth's surface to the geoid, by means of the free air gradient, gives approximately the value we would get in the absence of topography (and thus also in the absence of isostatic compensation), i.e., the value we are interested in. A geoid computed from free air anomalies, called in brief the *free air geoid*, shows little correlation with topography—see, for instance, FIG. 5 [VINCENT ET AL., 1972]. This is what one should expect under the hypothesis of isostasy, and the appropriateness of free air anomalies for this task is thus confirmed experimentally as well.

Another possibility is to use the isostatically compensated anomalies discussed in §21.4. The argument for using these goes as follows: Since the gravity field at the level of the geoid should behave much as if there were no masses above it and as if the masses underneath were distributed regularly, then we can say that if gravity is properly reduced to the geoid, i.e., taking both the topography and isostasy into

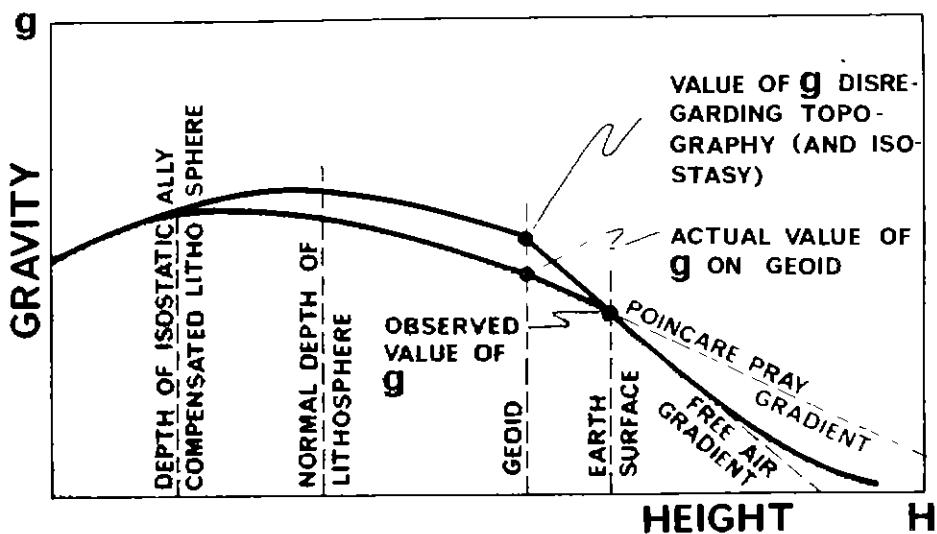


FIG. 22.4. Variations of gravity in a real and an idealized earth. (Real earth shaded.)

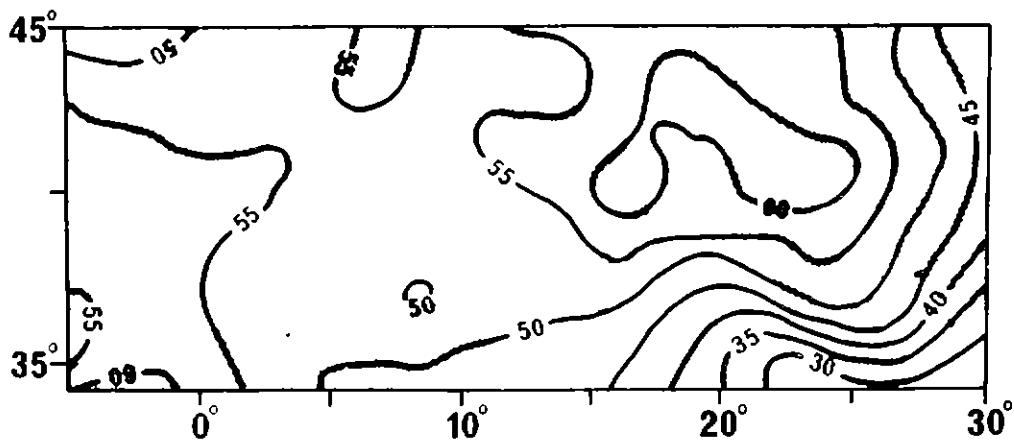


FIG. 22.5. Free air geoid. Contours in metres.

consideration, we do not have to worry about the redundant masses any more. This argument makes sense for the (about) 90% of the earth's surface that is in isostatic equilibrium. Discrepancies should be expected for the remaining 10 percent. To illustrate this, in FIG. 6 we show an *isostatically compensated geoid* computed by HEISKANEN [1957]. Allowing for different reference ellipsoids, one can see that the trends in both solutions follow the same broad pattern. The differences probably reflect the statistical uncertainty in both solutions more than the intrinsic difference in the two kinds of anomalies.

It has been pointed out that there are other possible ways of defining anomalies that satisfy the basic requirement of the geodetic boundary value problem. These, however, have not found widespread use. The interested reader is advised to pursue this matter in, e.g., HEISKANEN AND MORITZ [1967].

To conclude, let us turn our attention to the effect the 'squashing' (of undesired masses onto the geoid) has on the solution  $T$  and on the resulting geoid. This effect is known as the *indirect effect on the computed T or N*, and the geoid for which the indirect effect was not removed is called the *cogeoid*; thus we should really speak

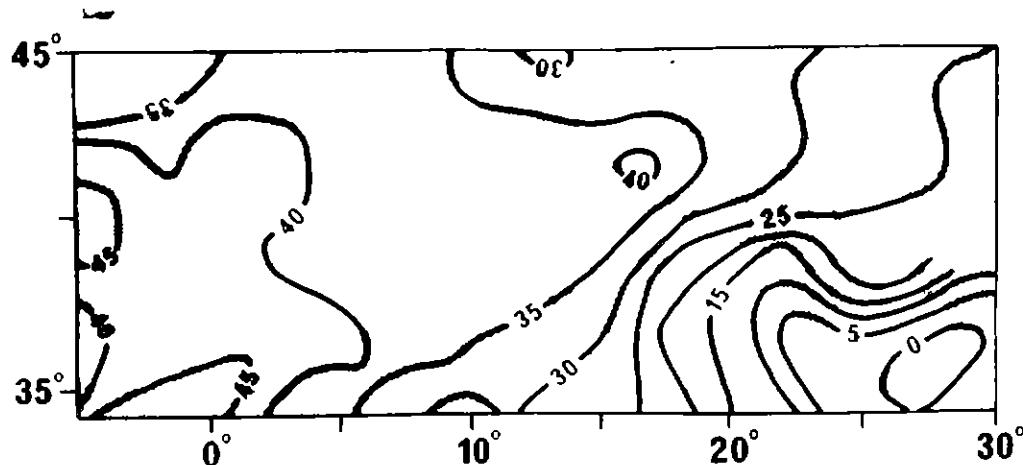


FIG. 22.6. Isostatically compensated geoid. Contours in metres.

about free air, or isostatically compensated, cogeoids.

The indirect effect on  $T$  may be evaluated as the difference,  $\delta W_M$ , between the potential of the neglected masses (to restore their effect on  $T$ ) and that,  $\delta W_\rho$ , of the introduced layer (to suppress its effect on  $T$ ). Using, for example, cylindrical coordinates, for the first potential we get the obvious equation (cf. FIG. 7):

$$\delta W_M(\phi_A, \lambda_A) = G\sigma \int_{\alpha=0}^{2\pi} \int_{r=0}^{\infty} \int_{z=0}^{H^1(r, \alpha)} \frac{r}{\rho'} dz dr d\alpha. \quad (22.29)$$

For the second potential, the following equation can be written in a similarly transparent way:

$$\delta W_\rho(\phi_A, \lambda_A) = G\sigma \int_{\alpha=0}^{2\pi} \int_{r=0}^{\infty} H^0(r, \alpha) dr d\alpha. \quad (22.30)$$

Evaluating the innermost integral in eqn. (29) (with respect to  $z$ ) and taking the difference of the two above expressions, we arrive at

$$\delta T^1(\phi_A, \lambda_A) = G\sigma \int_{\alpha=0}^{2\pi} \int_{r=0}^{\infty} \left\{ r \ln \left[ \frac{H^0(r, \alpha)}{r} + \sqrt{\left( 1 + \frac{(H^0(r, \alpha))^2}{r^2} \right)} \right] - H^0(r, \alpha) \right\} dr d\alpha. \quad (22.31)$$

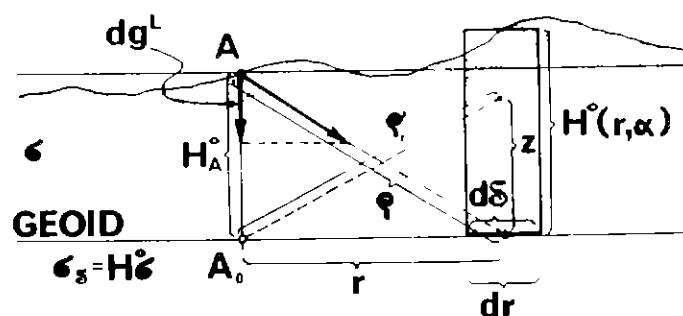


FIG. 22.7. Indirect effect of free air anomaly.

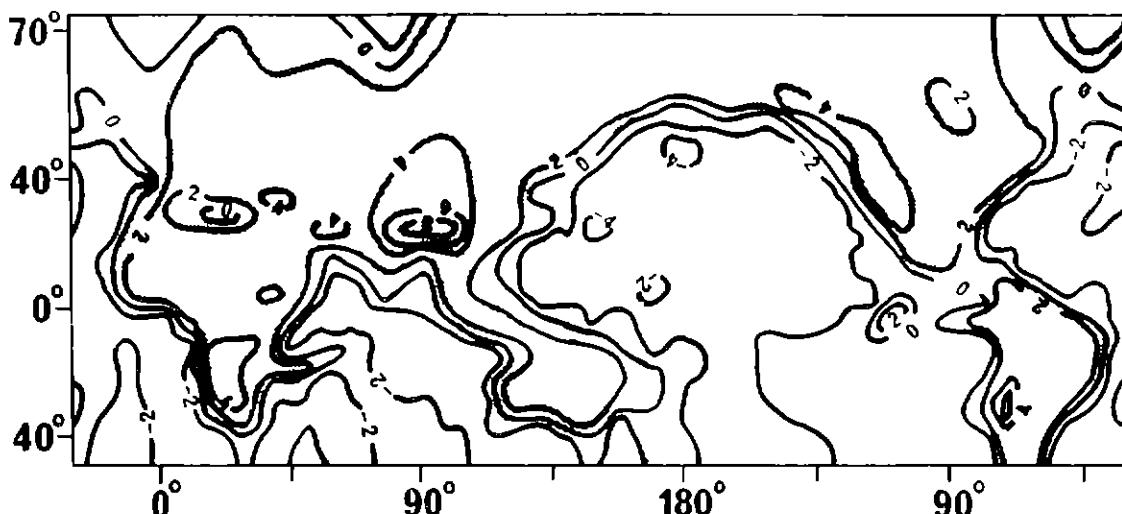


FIG. 22.8. Indirect effect on isostatically compensated anomalies. Contours in milligals.

The correction to cogeoidal height  $N$  is obtained from eqn. (31) through the application of the Bruns formula (eqn. (21.4)). It reaches the maximum of several metres, while the indirect effect of the isostatically compensated anomalies reaches as much as 10 m [HEISKANEN AND VENING MEINESZ, 1958].

Alternatively, the anomalies used in Stokes's formula for the cogeoid may be corrected directly beforehand. For illustration, HEISKANEN AND NISKANEN's [1941] results for the indirect effect on isostatically compensated anomalies are shown in FIG. 8.

One can also speak, formally, of the Bouguer or Poincaré–Pray cogeoids. These would simply be evaluated from the Bouguer or Poincaré–Pray (see §21.2) gravity anomalies through Stokes's formula without regard to their physical meaning. They would, however, deviate from the geoid so much as to make the cogeoid unrecognizable. According to HEISKANEN AND VENING MEINESZ [1958], the indirect effect of the Bouguer anomaly can reach more than 500 m, i.e., several times more than the geoidal height itself. The Poincaré–Pray cogeoid would appear to be even worse. The reason for such a huge indirect effect of these two kinds of anomalies is that they obviously violate the basic assumption behind the geodetic boundary value problem of no masses outside the geoid.

In practice, neither Stokes's nor Vening Meinesz's formulae are evaluated through the surface integration. Because of the inherent discreteness of measured gravity and thus of the discreteness of gravity anomalies, the surface integrals are replaced by summations. These summations may be considered as numerical techniques for evaluating the surface integrals, and they will be discussed in §22.4.

## 22.2. Molodenskij's concept

An alternative to Stokes's (Green's) solution to the geodetic boundary value problem is the approach by Molodenskij that employs integral equations. Once

again, the solution in terms of the disturbing potential  $T$  is sought. The disturbing potential is then transformed into the height anomaly  $\xi$  and the Molodenskij deflection of the vertical  $\tilde{\theta}$  (cf. §21.1). The main difference from the Stokes approach is that the problem is formulated for the surface of the earth rather than the geoid. The fact that the problem is formulated for the earth's surface is the main advantage of this approach—that of being able to solve for  $T$  without having to rely on any hypotheses concerning the mass distribution within the earth.

To begin with, let us recall (20.98) for the disturbing potential. Evidently, this equation could be converted into an equation for the height anomaly  $\xi$  by means of (21.10), and then one could attempt to solve for  $\xi$ . It is more expedient, however, to first solve for  $T$  and then convert it to  $\xi$  later. To solve for  $T$ , let us introduce the first of a series of required approximations: let us assume that (20.98) is valid on the telluroid rather than on the earth's surface. Denoting the telluroid by  $S'$ , we have

$$T(\bar{r}_A) = \frac{1}{2\pi} \iint_{S'} \left( T(\bar{r}) \frac{\partial \rho^{-1}}{\partial n'} - \frac{1}{\rho} \frac{\partial T(\bar{r})}{\partial n'} \right) dS', \quad (22.32)$$

where  $n'$  is the outer normal to the telluroid. This approximation is admissible because the subintegral function is small. It is equivalent to that done in §22.1 with Stokes's theory when the indirect effect was first neglected. Since the telluroid departs from the earth's surface by about the same amount as the geoid departs from the ellipsoid (cf. §7.4), the indirect effect and thus even the error of the approximation would be almost the same. Also note that the normal to the earth's surface  $n$  deviates from the normal to the telluroid  $n'$  by an amount equal to the slope of the quasigeoid (cf. FIG. 9), i.e., by the amount equivalent to Molodenskij's deflection of the vertical  $\tilde{\theta}$ .

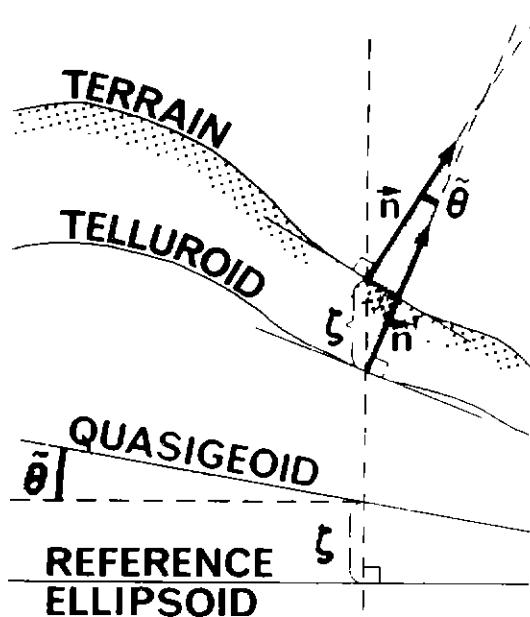


FIG. 22.9. Normals to the terrain and the telluroid.

Let us now turn to (32). MOLODENSKIY ET AL. [1960] have shown that the normal derivative of  $T$  can be closely approximated by

$$\frac{\partial T}{\partial n'} \doteq \left[ \frac{\partial T}{\partial H^N} + \gamma(\xi \tan \beta_1 + \eta \tan \beta_2) \right] \cos \beta, \quad (22.33)$$

where  $\gamma$  and  $\partial T / \partial H^N$  refer to the telluroid ( $H^N$ , the normal height—cf. §16.4—is being reckoned in the direction of the normal plumb line);  $\xi, \eta$  are the Molodenskij deflection components;  $\beta_1, \beta_2$  are the slopes of the north-south and east-west telluroid profiles; and  $\beta$  is the maximum slope of the telluroid at the point of evaluation. Expressing  $\partial T / \partial H^N$  from (21.16), i.e., the Molodenskij version of the fundamental gravimetric equation, and substituting back into (32), we obtain

$$\begin{aligned} T - \frac{1}{2\pi} \oint_{S'} & \left( \frac{\partial \rho^{-1}}{\partial n'} - \frac{\cos \beta}{\rho \gamma} \frac{\partial \gamma}{\partial H^N} \right) T dS' = \\ &= \frac{1}{2\pi} \oint_{S'} \left( \frac{\cos \beta}{\rho} \widetilde{\Delta g} - \frac{\gamma \cos \beta}{\rho} (\xi \tan \beta_1 + \eta \tan \beta_2) \right) dS'. \end{aligned} \quad (22.34)$$

This is an integral equation of the Fredholm type (§3.2) for  $T$ , where all the quantities are observable to a certain degree of accuracy except for  $T$ . Therefore, theoretically, it could be solved for  $T$  if  $\beta_1, \beta_2, \beta, \widetilde{\Delta g}, \xi, \eta$  were known all over the earth's surface. However, the equation is rather cumbersome and some simplifications are usually sought.

One such simplification can be achieved through the use of the surface density function (see §20.4). Let us express the disturbing potential on the telluroid by means of (20.101): we get

$$T(\bar{r}_A) = \oint_{S'} \frac{\Phi}{\rho} dS'. \quad (22.35)$$

Its derivative with respect to the normal plumb line at  $A$  is

$$\frac{\partial T}{\partial H^N} \Big|_A = \oint_{S'} \frac{1}{\rho} \frac{\partial \Phi}{\partial H^N} \Big|_A dS' + \oint_{S'} \Phi \frac{\partial \rho^{-1}}{\partial H^N} \Big|_A dS', \quad (22.36)$$

where the first term can be shown to be equal to [MOLODENSKIY ET AL., 1960]

$$\oint_{S'} \frac{1}{\rho} \frac{\partial \Phi}{\partial H^N} \Big|_A dS' = \begin{cases} 0 & \text{for } A \text{ not on the telluroid,} \\ -2\pi\Phi \cos \beta & \text{for } A \text{ on outer side of telluroid,} \\ 2\pi\Phi \cos \beta & \text{for } A \text{ on inner side of telluroid.} \end{cases} \quad (22.37)$$

Limiting ourselves to the outer side of the telluroid, we may now substitute (35) and (36) into (21.16) and obtain, after a trivial rearrangement,

$$\Phi(\bar{r}_A) - \frac{1}{2\pi \cos \beta_A} \oint_{S'} \left( \frac{\partial \rho^{-1}}{\partial H^N} - \frac{1}{\gamma \rho} \frac{\partial \gamma}{\partial H^N} \right) \Phi dS' \doteq \frac{\widetilde{\Delta g}_A}{2\pi \cos \beta_A}. \quad (22.38)$$

Further simplification may be achieved by expressing  $\rho$  as a function of the positions of the computation and dummy points from (20.49). As stated in §20.3, the deviation between the ellipsoidal normal and the radius vector is never larger than 13 minutes of arc. Therefore, even the deviation  $\delta$  between the normal plumb line and the radius vector  $r_A$  is of the same order—see FIG. 10. Hence, to an accuracy better than  $e^2$ , the derivative with respect to  $H^N$  can be replaced by a derivative with respect to  $r_A$ . After a few operations, we obtain

$$\frac{\partial \rho^{-1}}{\partial H^N} \doteq -\frac{r_A - r \cos \psi}{\rho^3}. \quad (22.39)$$

On the other hand, taking  $\partial \gamma / \partial H^N$  from (21.30), the second term under the integration sign in (38) becomes

$$\frac{1}{\gamma \rho} \frac{\partial \gamma}{\partial H^N} \doteq -\frac{2\gamma_0}{\gamma \rho (a + H^N)} (1 + m + f \cos 2\phi) \doteq -\frac{2}{\rho r_A}. \quad (22.40)$$

Thus the kernel can be rewritten as follows:

$$K(\vec{r}_A, \vec{r}) = \frac{\partial \rho^{-1}}{\partial H^N} - \frac{1}{\gamma \rho} \frac{\partial \gamma}{\partial H^N} \doteq \frac{2}{\rho r_A} - \frac{r_A - r \cos \psi}{\rho^3}. \quad (22.41)$$

Employing (20.49) again, we get

$$K(\vec{r}_A, \vec{r}) \doteq \frac{3}{2r_A \rho} - \frac{r_A^2 - r^2}{2r_A \rho^3}. \quad (22.42)$$

Lastly, the form of the kernel can be made more convenient by some spherical approximations. The telluroid surface element  $dS'$  can be written as (cf. FIG. 11)

$$dS' = \frac{r^2}{\cos \beta} d\nu, \quad (22.43)$$

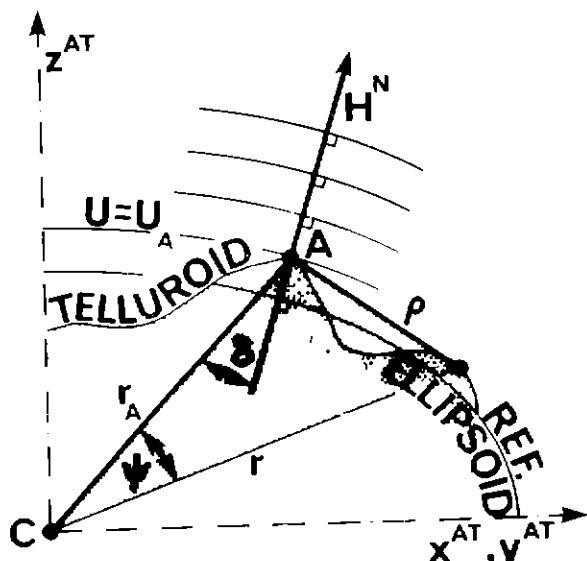


FIG. 22.10. Approximate expression for the Molodenskij kernel.

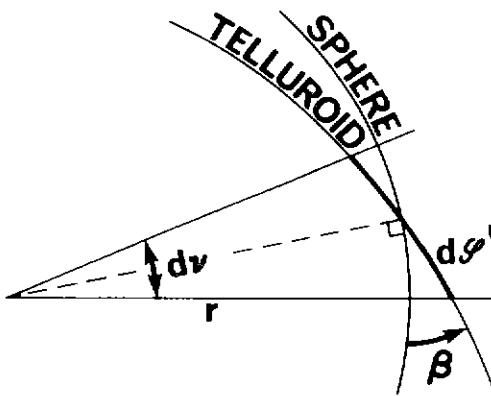


FIG. 22.11. Solid angle and surface elements.

where  $d\nu$  is the solid angle element. With the same degree of approximation, i.e.,  $e^2$ , we can write

$$r_A \doteq R + H_A^N, \quad r \doteq R + H^N, \quad (22.44)$$

where  $R$  is once more  $\sqrt[3]{a^2 b}$ , and  $H^N$  are (normal) heights of the telluroid above the ellipsoid. By denoting  $\Phi/\cos \beta$  by  $\Phi'$ , the final form of (38) becomes

$$\begin{aligned} \Phi'(\bar{r}_A) - \frac{R}{4\pi \cos^2 \beta_A} \iint_{S'} \left( \frac{3}{\rho_0} + \frac{2R(H^N - H_A^N)}{\rho_0^3} \right. \\ \left. - \frac{3(H^N - H_A^N)^2}{2\rho_0^3} + \dots \right) \Phi' d\nu \doteq \frac{\widetilde{\Delta g}_A}{2\pi \cos^2 \beta_A}, \end{aligned} \quad (22.45)$$

where  $\rho_0 \doteq \{\rho^2 - (H^N - H_A^N)^2\}^{1/2}$  is the chord distance between the projections of points  $\bar{r}_A, \bar{r}$  onto the quasigeoid.

The solution of (45) is usually attempted in an iterative manner (see §3.2). One notices that the first term under the integral sign is, except for the immediate neighbourhood of  $A$ , larger than the remaining two terms. Thus the integral over the first term will be generally larger than the integrals of the second and third terms. Also, if the terrain around  $A$  is flat,  $\cos^2 \beta_A$  reduces to 1. It is thus convenient to specify the 0th iteration,  $\Phi'^{(0)}$ , as involving only the first term and, in the beginning, assume the terrain to be flat. One thus has

$$\Phi'^{(0)}(\bar{r}_A) = \frac{\widetilde{\Delta g}_A}{2\pi} + \frac{3R}{4\pi} \iint_{S'} \frac{\Phi'^{(0)}(\bar{r})}{\rho_0(\bar{r}_A, \bar{r})} d\nu. \quad (22.46)$$

It can be shown, using (35) and (16) formulated on the telluroid, that

$$\iint_{S'} \frac{\Phi'^{(0)}}{\rho_0} d\nu \doteq \frac{1}{4\pi R} \iint_{S'} \widetilde{\Delta g} S(\psi) d\nu, \quad (22.47)$$

and the 0th iteration becomes

$$\Phi'^{(0)}(\bar{r}_A) \doteq \frac{\widetilde{\Delta g}_A}{2\pi} + \frac{3}{16\pi^2} \iint_{S'} \widetilde{\Delta g} S(\psi) d\nu. \quad (22.48)$$

The first iteration  $\Phi'^{(1)}$  for a flat terrain can then be evaluated from (45) taking the first two subintegral terms. Denoting  $\Phi'^{(1)} = \Phi'^{(0)} + \delta\Phi'^{(1)}$  and neglecting the product of the second term with  $\delta\Phi'^{(1)}$ , we get

$$\begin{aligned} \Phi'^{(1)}(\bar{r}_A) &\doteq \frac{\widetilde{\Delta g}_A}{2\pi} + \frac{R}{4\pi} \left( 3 \iint_{S'} \frac{\Phi'^{(0)}}{\rho_0} d\nu + 2R \iint_{S'} \frac{H^N - H_A^N}{\rho_0^3} \Phi'^{(0)} d\nu \right. \\ &\quad \left. + 3 \iint_{S'} \frac{\delta\Phi'^{(1)}}{\rho_0} d\nu \right). \end{aligned} \quad (22.49)$$

Subtracting the 0th iteration and denoting

$$\Delta g_A^{(1)} = R^2 \iint_{S'} \frac{H^N - H_A^N}{\rho_0^3} \Phi'^{(0)} d\nu, \quad (22.50)$$

we obtain the integral equation for  $\delta\Phi'^{(1)}$  in the following form:

$$\delta\Phi'^{(1)}(\bar{r}_A) = \frac{\Delta g_A^{(1)}}{2\pi} + \frac{3R}{4\pi} \iint_{S'} \frac{\delta\Phi'^{(1)}}{\rho_0} d\nu. \quad (22.51)$$

Since the form of (51) is identical with that of (46), except that  $\widetilde{\Delta g}$  is replaced by  $\Delta g^{(1)}$ , the solution  $\delta\Phi'^{(1)}$  must then be given by an equation equivalent to (48); namely,

$$\delta\Phi'^{(1)}(\bar{r}_A) = \frac{\Delta g_A^{(1)}}{2\pi} + \frac{3}{16\pi^2} \iint_{S'} \Delta g^{(1)} S(\psi) d\nu. \quad (22.52)$$

The first iteration then becomes

$$\boxed{\Phi'^{(1)}(\bar{r}_A) \doteq \frac{\widetilde{\Delta g}_A + \Delta g_A^{(1)}}{2\pi \cos^2 \beta_A} + \frac{3}{16\pi^2 \cos^2 \beta_A} \iint_{S'} (\widetilde{\Delta g} + \Delta g^{(1)}) S(\psi) d\nu, \quad (22.53)}$$

as the reader can verify. This is the integral relation between the surface density function and the surface gravity anomaly.

Higher order approximations can be derived following a similar argument. However, since the third subintegral term in (45) is much smaller than the first two, the first approximation is adequate to the relative accuracy of  $e^2$ . If a solution in high mountains is sought, it may be necessary to take the third term into consideration [MOLODENSKIJ ET AL., 1960].

Let us now try to get the solution for the height anomaly  $\zeta$ . Recalling (21.10) and combining it with (35), one gets

$$\zeta = \frac{1}{\gamma} \iint_{S'} \frac{\Phi}{\rho} dS' \doteq \frac{R^2}{\gamma} \iint_{S'} \frac{\Phi'}{\rho_0} d\nu. \quad (22.54)$$

When only the first approximation  $\xi^{(1)}$  is sought, the first approximation of  $\Phi'$ , i.e.,  $\Phi'^{(1)}$ , may be enough. Then (47) can be used for  $\Phi'^{(0)}$  and a similar equation for  $\delta\Phi'^{(1)}$  to give us

$$\xi^{(1)} = \xi^{(0)} + \delta\xi^{(1)} \doteq \frac{R}{4\pi\gamma_0} \iint_{S'} (\widetilde{\Delta g} + \Delta g^{(1)}) S(\psi) d\nu, \quad (22.55)$$

where normal gravity  $\gamma_0$  on the ellipsoid was substituted for  $\gamma$ .

Note that  $\Delta g^{(1)}$  can be regarded as showing the effect of topography ( $H^N - H_A^N$ ) on the surface gravity anomalies. For a more detailed discussion of this effect and its relation to the topographical correction (cf. §21.4) see MORITZ [1968].

Clearly,  $\delta\xi^{(1)}$  can also be regarded as the effect of topography on the height anomaly. Since the geoid, and thus the quasigeoid as well, are little affected by topography,  $\delta\xi^{(1)}$  must be considered as showing the inherent effect of topography on the surface anomaly  $\widetilde{\Delta g}$ . Is there then a way of evaluating  $\Delta g^{(1)}$  without the necessity of computing the surface density function? Yes. We may express  $\Phi'^{(0)}$  in (50) using (48), where the integral is a linear function of  $\xi^{(0)}$ , and obtain

$$\boxed{\Delta g_A^{(1)} = \frac{R^2}{2\pi} \iint_{S'} \frac{H^N - H_A^N}{\rho_0^3} \left( \widetilde{\Delta g} + \frac{3\gamma}{2R} \xi^{(0)} \right) d\nu,} \quad (22.56)$$

which is a much more convenient formula. In most cases, it suffices to take just the first term under the integration sign [HEISKANEN AND MORITZ 1967]. Evidently, the subintegral function tapers off very rapidly because of the fast growing denominator, so the integration does not have to be carried out very far beyond the immediate neighbourhood of  $A$ .

Let us note that (55) is the Molodenskij equivalent to Stokes's formula (17) and requires a reference ellipsoid with properly assessed mass. If this is not the case, an absolute term  $\xi_0$  similar to  $N_0$  (cf. (18)) must be added to the solution. The equivalents of Vening Meinesz's formulae for the geoidal deflection components also exist in the Molodenskij approach. The so-determined deflection components are, however, of the Molodenskij kind (cf. §21.1). We quote MOLODENSKIY ET AL. [1960] for the final result:

$$\boxed{\begin{Bmatrix} \xi \\ \eta \end{Bmatrix} \doteq \frac{1}{4\pi\gamma_0} \iint_{S'} (\widetilde{\Delta g} + \Delta g^{(1)}) \begin{Bmatrix} \cos\alpha \\ \sin\alpha \end{Bmatrix} \frac{dS(\psi)}{d\psi} d\nu - \frac{\widetilde{\Delta g}}{\gamma_0} \begin{Bmatrix} \tan\beta_1 \\ \tan\beta_2 \end{Bmatrix}} \quad (22.57)$$

It is of interest now to compare the geoid and the quasigeoid derived from the same observed gravity data on the surface of the earth. Taking (17) and (55) and substituting for the two kinds of gravity anomalies (using (21.30) and (21.32)), we

obtain

$$N - \xi \doteq \frac{R}{4\pi\gamma_0} \oint_{S'} \left( 0.3086 H^O - \frac{2\gamma_0}{a} (1 + m + 2f \cos^2 \phi) H^N - \Delta g^{(1)} \right) S(\psi) d\nu, \quad (22.58)$$

where  $H^O$  is the orthometric and  $H^N$  the normal heights. Substituting  $H^O \bar{g}' / \bar{\gamma}$  for  $H^N$  from (16.97) and (16.100), we get

$$N - \xi \doteq - \frac{R}{4\pi\gamma_0} \oint_{S'} \left( 0.3086 \frac{H^O \Delta g^{(B)}}{\gamma_0} + \Delta g^{(1)} \right) S(\psi) d\nu, \quad (22.59)$$

for  $H^O$  in metres;  $\gamma_0, \Delta g^{(1)}$  in milligals; and  $\Delta g^{(B)}$ , the Bouguer anomaly, also in milligals. The high correlation of this difference with topography is clearly discernible. The difference can reach a few metres in the mountains, as already stated in §7.4. Note that (59) allows one to compute either the geoidal or quasigeoidal height from any of the appropriate anomalies if the anomalies are properly corrected for topographical effect. This correction comes from (56).

A similar comparison of the geoidal and Molodenskij deflection components yields

$$\begin{aligned} \theta - \tilde{\theta} = & \left\{ \begin{array}{l} \xi \\ \eta \end{array} \right\} - \left\{ \begin{array}{l} \tilde{\xi} \\ \tilde{\eta} \end{array} \right\} \doteq - \frac{1}{4\pi\gamma_0} \oint_{S'} \left( 0.3086 \frac{H^O \Delta g^{(B)}}{\gamma_0} + \Delta g^{(1)} \right) \\ & \times \left\{ \begin{array}{l} \cos \alpha \\ \sin \alpha \end{array} \right\} \frac{dS(\psi)}{d\psi} d\nu + \frac{\widetilde{\Delta g}}{\gamma_0} \left\{ \begin{array}{l} \tan \beta_1 \\ \tan \beta_2 \end{array} \right\}. \end{aligned} \quad (22.60)$$

This formula can be used in determining the effect of the curvature of the actual plumb line on the deflection components. As was seen in §21.1, if all three kinds of deflections are referred to the same reference ellipsoid, the difference between the Molodenskij and surface deflections is given by the amount of curvature of the normal plumb line between the ellipsoid and the telluroid (cf. (21.59) and (21.62)). Then the difference between the surface and geoidal deflections is governed by the curvature of the actual plumb line (cf. §21.3). Hence  $\theta - \tilde{\theta}$ , after being properly corrected for normal curvature, can give us the amount of actual curvature. This situation is summarized in FIG. 12. Thus (59) and (60) provide us with the vital links between Stokes's and Molodenskij's approaches. These links clearly involve the topographical and curvature effects.

There exist other approaches to the determination of the field parameters from gravity. These have not seen very wide applications yet, and so they will not be dealt with here. The interested reader is advised to pursue this topic in, e.g., HIRVONEN [1960], BJERHAMMAR [1963], KRARUP [1973], and MORITZ [1979].

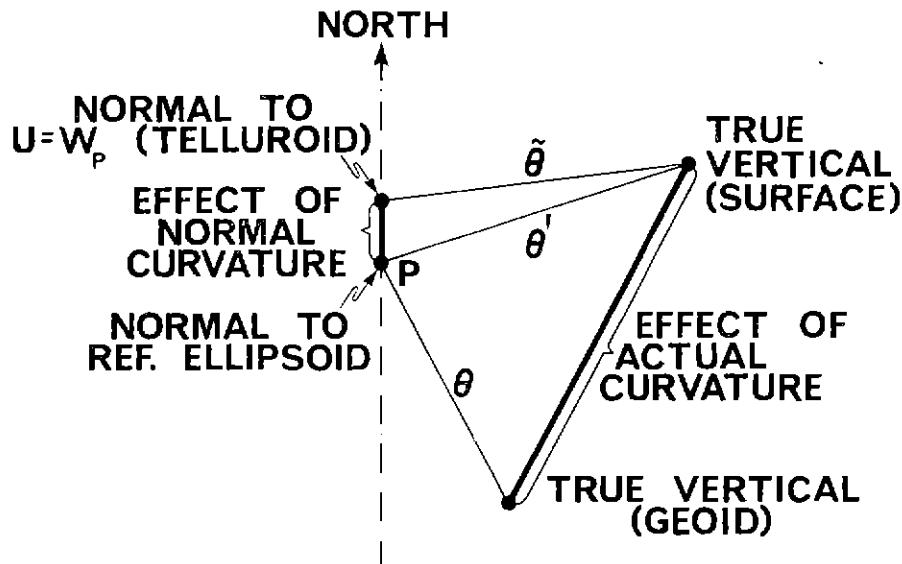


FIG. 22.12. Deflections and curvature effects (plotted on tangent plane to the reference ellipsoid at point of interest  $P$ ).

### 22.3. Gravimetry

The previous two sections showed how gravity magnitude observations are converted into other parameters of the earth's gravity field. So far, nothing has been said about how the gravity data is acquired and processed to obtain a consistent set of global anomalies appropriate for use in either of the two approaches described above.

Gravity (acceleration) at a point on the surface of the earth can be determined directly by either *pendulums* or *free-fall devices*. For the instrumental aspects, the interested reader is referred to either standard textbooks or special papers such as FALLER [1965], VALLIANT [1971], or TELFORD ET AL. [1976]. Suffice it to say that, in both cases, precise timing of movements is required, and the observed time interval is subsequently converted into the value of  $g$  using the appropriate equation of motion. The achievable accuracy, depending largely on control of the instrument environment and the number of repeated observations, is of the order of or better than  $100 \mu\text{Gal}$  for the pendulums and perhaps one order of magnitude better for the free-fall devices. Both kinds of instruments give us *absolute gravity measurements*.

More portable, handier to operate, and thus more widely used are *gravimeters*, instruments capable of measuring only differences in gravity between pairs of points. There is a multitude of designs for which the reader is referred to, e.g., MORELLI [1963]. Since gravimeters give us only *relative gravity measurements*, their accuracy is usually significantly higher than that of the absolute instruments. The accuracy routinely achieved is about  $50 \mu\text{Gal}$ , but instruments accurate to a fraction of a  $\mu\text{Gal}$  exist [GOODKIND, 1978]. Inertial positioning devices, operated in a special mode, also have the capability of measuring local gravity variations between two points where the gravity values are known. The accuracy of gravity predicted in this way is about  $\sigma_g = 3 \text{ mGal}$  [GREGERSON, 1979].

In practice, one tries to combine advantages of both techniques. This is done through the establishment of networks of gravity stations using both absolute and relative measurements. The stations with absolute gravity determination provide the anchoring points of the network, while the relative measurements provide the ties between the points. The establishment of these *global gravity networks* has been coordinated internationally since the beginning of this century. As early as 1909, the IAG adopted the international *Potsdam Gravity System*, a network used for a multitude of tasks. This system has only recently been replaced by the *International Gravity Standardization Net 1971* (IGSN 71) [IAG, 1974] shown in FIG. 13.

When the absolute and relative observations are made and assessed for accuracy, an adjustment can be carried out using any of the adjustment techniques discussed in Part III. The adjustment results in the estimated values of gravity for all the stations, together with their accuracy estimates. In the IGSN 71, the accuracy of gravity at any station is on average 36  $\mu\text{Gal}$  [IAG, 1974]. The adjustment procedure is practically identical with that of geodetic levelling (cf. §19.2).

Once the global network is developed, it can be densified by acquiring new, usually relative, measurements. This is normally done by individual countries through the development of *national gravity networks* adjusted to fit within the global framework. The detailed gravity points are then tied to the network by adjusting the individual gravimetric legs (traverses) into the national framework. An example of a national coverage is shown for Canada in FIG. 14.

The detailed points are observed on land, on the sea surface, on the sea beds, and on the bottoms of lakes. Needless to say, the determination of coordinates of the individual points is an indispensable part of the data acquisition process. Of these coordinates, the height is the most crucial because of the required higher accuracy in the vertical sense. There is a much stronger dependency of  $g$  on height than on

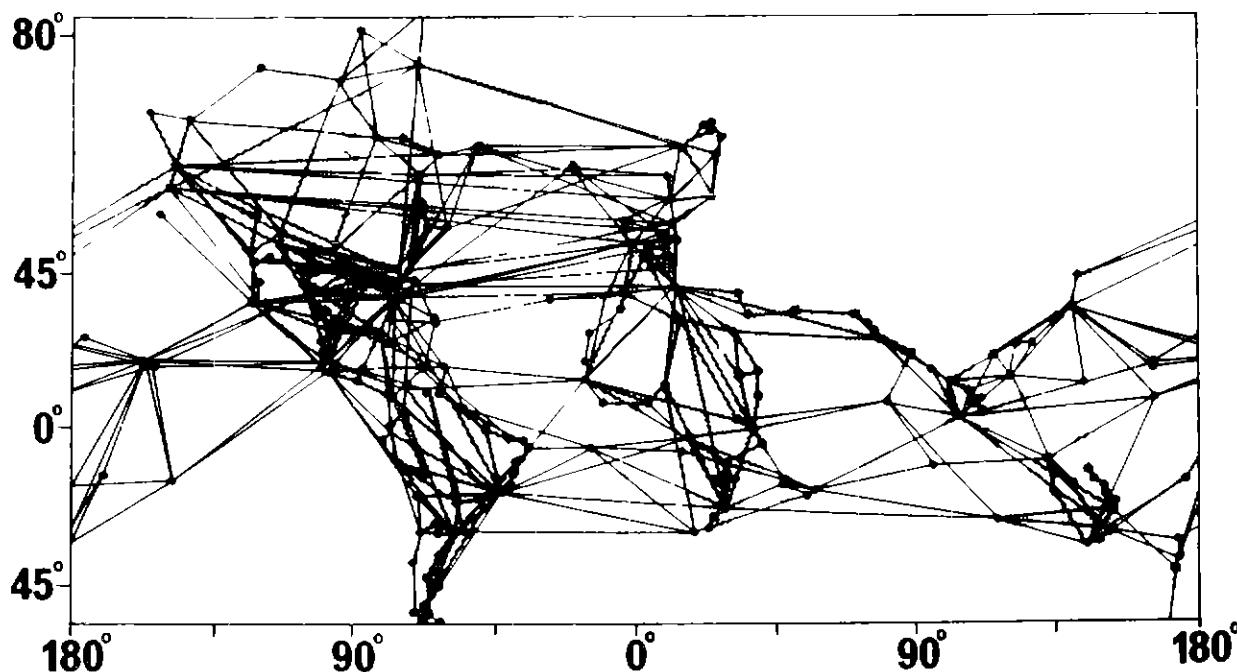


FIG. 22.13. International Gravity Standardization Net 1971.

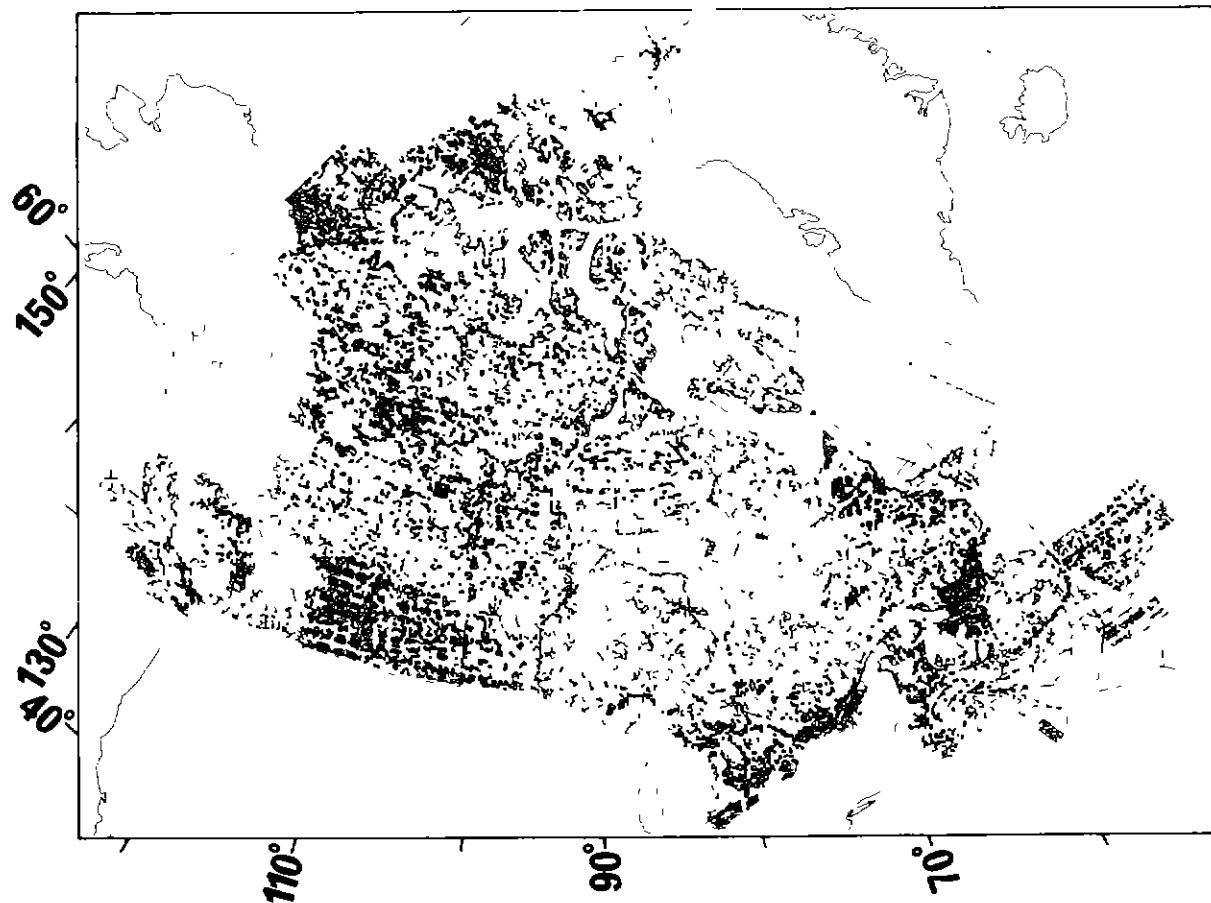


FIG. 22.14. Gravity data coverage in Canada. (Courtesy of Earth Physics Branch, DEPARTMENT OF ENERGY, MINES AND RESOURCES [1977], Ottawa, Canada).

horizontal position: while 50 to 200 metre accuracy in a horizontal position is considered adequate, height should be determined as accurately as practicable. In practice, the height accuracy varies from a few decimetres, when heights are determined by levelling, to a few metres for barometric heighting.

After the detailed point gravity values have been adjusted, the *point anomalies* can be evaluated. For this purpose, the surface gravity value  $g$  must be reduced to the geoid ( $g_0$ ) using the appropriate gravity gradient (cf. §6.2):

$$g_0 = g + \frac{\partial g}{\partial H} H^O. \quad (22.61)$$

In other words, the surface gravity must be corrected for the effect of the height  $H^O$  of the observing station. Clearly, every error made in the determination of the station height is transformed by the vertical gradient of gravity into an error in anomaly. Using the law of propagation of errors (see (11.20)), one obtains

$$\sigma_{\Delta g}^2 = \sigma_g^2 + \left( \frac{\partial g}{\partial H} \right)^2 \sigma_H^2, \quad (22.62)$$

whereby  $\sigma$  denotes the standard deviations in the quantities involved. Considering

the free air anomaly, for example, it is not difficult to calculate that an error of 1 m in height contributes six times more to the error in the anomaly than the observation error in  $g$  (normally equal to  $50 \mu\text{Gal}$ ) does.

The computed point anomalies are then stored; in this form they are used for various geophysical and geodetic investigations. For geodetic applications, however, a global coverage that is as homogeneous as possible is also needed (see §22.1 and §22.2). For this reason, global anomaly files are used where mean anomalies are assigned to surface elements (cells) of a selected size delineated ordinarily by geodetic coordinates  $\phi$  and  $\lambda$ . These cells normally have one of the following sizes:  $5' \times 5'$ ,  $20' \times 20'$ ,  $1^\circ \times 1^\circ$ ,  $5^\circ \times 5^\circ$ , or  $10^\circ \times 10^\circ$ . The availability of mean anomalies for the whole earth is illustrated in FIG. 15 [RAPP, 1977].

The mean anomalies  $\bar{\Delta}g$  are evaluated from point anomalies  $\Delta g$  within each cell using any one of the existing different techniques. Here, the integral formula is given as an example:

$$\bar{\Delta}g = \frac{1}{D} \iint_{\mathcal{D}} \Delta g d\mathcal{D}, \quad (22.63)$$

where  $D$  is the area of the cell  $\mathcal{D}$ . The interested reader can find other formulae in, e.g., MORITZ [1963] or RAPP [1964]. Standard deviations of mean anomalies are evaluated from the point anomaly standard deviations. They also depend on the distribution of point anomalies within the cell and are usually supplied, together with the mean anomalies.

The main problem with global gravity files is the inhomogeneity of the data coverage. Depending on the size of the cell, a large percentage of the surface elements do not contain any gravity observations at all. This situation occurs very

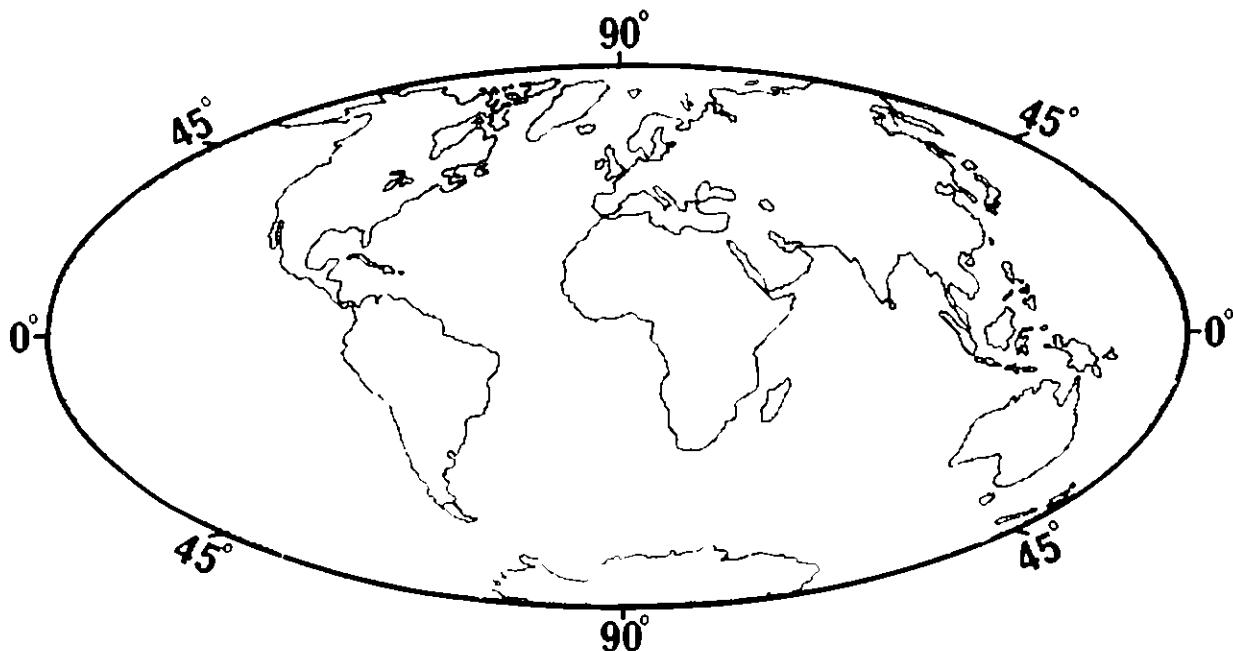


FIG. 22.15. Available mean  $1^\circ \times 1^\circ$  anomalies.

frequently on the seas, but there are also large hinterland areas, usually uninhabited, that suffer the same fate.

When relatively isolated empty cells are encountered, the task of *prediction of mean anomalies* for them is relatively straightforward. It involves bridging the gaps between neighbouring cells, which can be done using either regression (see §14.2) or collocation (see §14.3). The mathematical model for the regression is simply

$$\overline{\Delta g}(\phi, \lambda) = \tilde{\Phi}^T(\phi, \lambda) \mathbf{c}, \quad (22.64)$$

where  $\overline{\Delta g}$  denotes the known, as well as predicted, mean anomalies;  $\tilde{\Phi}^T(\phi, \lambda)$  is a column of Vandermonde's matrix composed of selected base functions (cf. §3.1); and  $\mathbf{c}$  is the vector of coefficients to be determined from the known anomalies. The base functions are selected arbitrarily; the simplest choice is the algebraic functions  $1, x, y, xy, x^2, \dots$ , with  $x, y$  indicating the local Cartesian coordinates

$$x = R(\phi - \phi_0), \quad y = R \cos \phi_0 (\lambda - \lambda_0), \quad (22.65)$$

and  $(\phi_0, \lambda_0)$  being a point preferably close to the centroid of the area of interest. (Note that this is an approximation to an LG system (cf. FIG. 15.23) with the direction of the  $y$ -axis reversed.) If the number of base functions, and thus the dimensions of the  $\mathbf{c}$  vector, is smaller than the number of surrounding known mean anomalies, we are faced with an overdetermined problem solvable by means of, e.g., the least-squares method. The standard deviations of the known anomalies are then used to construct the weight matrix in the usual fashion.

The other alternative is to regard the mean anomalies as random quantities with a zero mean. The known anomalies then become a two-dimensional random data series (with argument  $(\phi, \lambda)$ ) that can be decomposed into (cf. §14.3)

$$\overline{\Delta g}(\phi, \lambda) = s(\phi, \lambda) + v(\phi, \lambda), \quad (22.66)$$

where  $s$  is the statistically dependent part of  $\Delta g$  and  $v$  is not. The statistically dependent part can be predicted using (14.82) if the covariance function,

$$C(\overline{\Delta g}(\phi_i, \lambda_i), \overline{\Delta g}(\phi_j, \lambda_j)), \quad (22.67)$$

is known. HIRVONEN [1962] suggested that for distances under 100 km, a homogeneous and isotropic covariance function depending only on the distance  $d$  of any two anomalies may be used, namely,

$$C(d) = \frac{337}{1 + (d/40)^2} \text{ mGal}^2, \quad (22.68)$$

where  $d$  is in kilometres. For the global covariance, the shape shown in FIG. 16 was derived by KAULA [1963]. This shape reflects the rule of thumb for the lower order harmonics of the gravity field already mentioned in §6.4. Alternative expressions may be found in TSCHERNING AND RAPP [1974]. The problem with the statistical approach is that, unless we deal with at least a significant part of the earth's surface, we cannot safely assume that  $E(\overline{\Delta g}) = 0$ , and this basic assumption behind the least-squares collocation (cf. §10.3 and §14.3) breaks down.

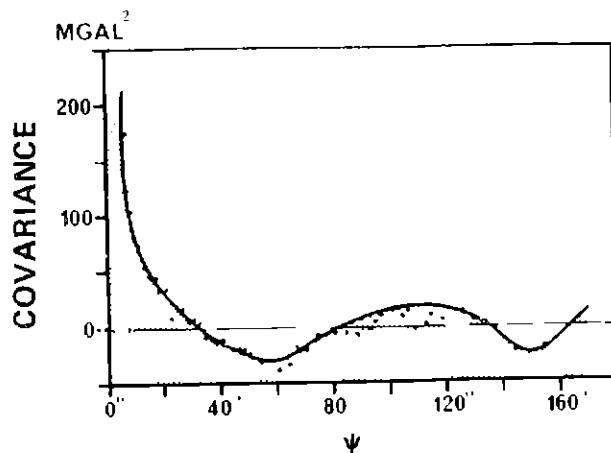


FIG. 22.16. Global correlation of gravity anomalies.

This problem can be alleviated by combining the collocation with the least-squares regression, where the former is used to further analyse the residuals from the latter. Another alternative is to first subtract from the existing anomalies the lower harmonic part, i.e., the harmonic series truncated at a certain order, and treat with collocation only the residual higher frequency content, which is more likely to have a zero mean [SCHWARZ AND LACHAPELLE, 1980].

It should be pointed out that, for the purpose of predicting new mean anomalies, the Bouguer anomaly is the most suitable. This is because it varies more smoothly than the other anomalies. Once a Bouguer mean anomaly has been predicted, it can, of course, be easily converted to any other kind of anomaly, if the appropriate height is known.

The situation becomes more serious when mean anomalies have to be predicted for very large areas where the known anomalies are far apart. In such cases, we have to either leave the cells empty or try to predict the mean values on the basis of our knowledge of the geology and physics of the earth's crust and upper mantle in that area. The interested reader may pursue these matters in, e.g., ORLIN [1966].

For completeness, it must be mentioned that for some applications it is advantageous to have the anomalies expressed globally in spherical harmonics rather than mean values. The way to do this was shown in §22.1. One such development has been carried out by RAPP [1977].

## 22.4. Evaluation of the surface integrals

In the course of determining the gravity field from gravity observations, we are faced with the necessity of evaluating the two kinds of convolution surface integrals:

$$\begin{aligned} & \oint_S f(\bar{r}) S(\psi(\bar{r}_A, \bar{r})) d\nu, \\ & \oint_S f(\bar{r}) \left\{ \begin{array}{l} \cos \alpha(\bar{r}_A, \bar{r}) \\ \sin \alpha(\bar{r}_A, \bar{r}) \end{array} \right\} \frac{dS(\psi)}{d\psi} d\nu, \end{aligned} \quad (22.69)$$

where  $f(\bar{r})$  is a finite function defined on the ellipsoid or telluroid  $\mathbb{S}$ . Once the values of the subintegral function are known for all the surface  $\mathbb{S}$ , the integration can be thought of as being carried out on the surface of a unit sphere.

The first point needing discussion is the *singularity of the surface integrals* for  $\psi = 0$ , i.e., for  $\bar{r} = \bar{r}_A$  (cf. §22.1). It can be shown that this singularity is removable: it either disappears with the convenient choice of a coordinate system, or it can be removed using some elementary mathematical tools. Let us start with the integrals possessing Stokes's kernel  $S(\psi)$ , and show how these can be evaluated when a *polar coordinate system*  $(\psi, \alpha)$  on the surface of the unit sphere is used—see FIG. 17. In this coordinate system, the solid angle element  $d\nu$  becomes (see §3.2)

$$d\nu = \sin \psi d\psi d\alpha, \quad (22.70)$$

and the surface integral can be written as

$$\iint_{\mathbb{S}} f(\bar{r}) S(\psi) d\nu = \int_{\alpha=0}^{2\pi} \int_{\psi=0}^{\pi} f(\bar{r}) S(\psi) \sin \psi d\psi d\alpha. \quad (22.71)$$

Denoting the new kernel  $S(\psi)\sin \psi$ , by  $F(\psi)$ , one obtains (cf. (15))

$$F(\psi) = 2 \cos \frac{1}{2}\psi - \sin \psi [6 \sin \frac{1}{2}\psi - 1 + \cos \psi (5 + 3 \ln (\sin \frac{1}{2}\psi + \sin^2 \frac{1}{2}\psi))]. \quad (22.72)$$

This is a very well behaved function with no singularities in  $\langle 0, \pi \rangle$ , and its graph is shown in FIG. 18. Then the expression  $\int_0^{2\pi} \int_0^\pi f(\psi, \alpha) F(\psi) d\psi d\alpha$  is easily integrated.

In some situations, the surface function  $f(\psi, \alpha)$ , be it a gravity anomaly or something else, can be regarded as a function of  $\psi$  only. This case is often met when one deals with idealized situations. Equation (71) then becomes

$$\iint_{\mathbb{S}} f(\bar{r}) S(\psi) d\nu = 2\pi \int_0^\pi f(\psi) F(\psi) d\psi. \quad (22.73)$$

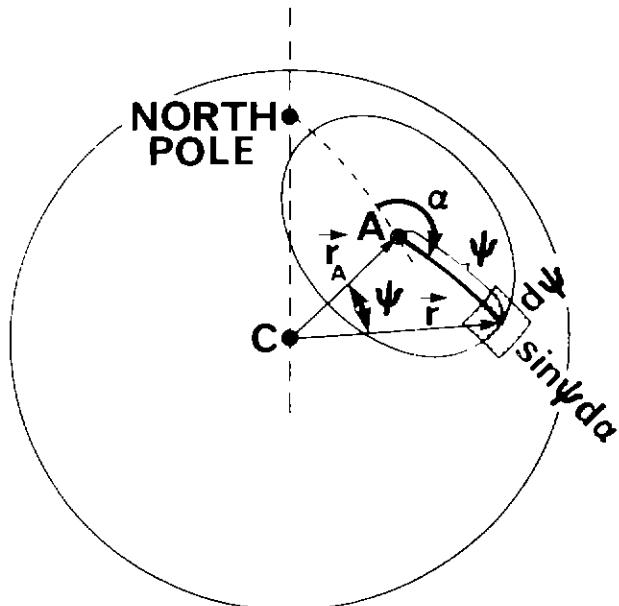
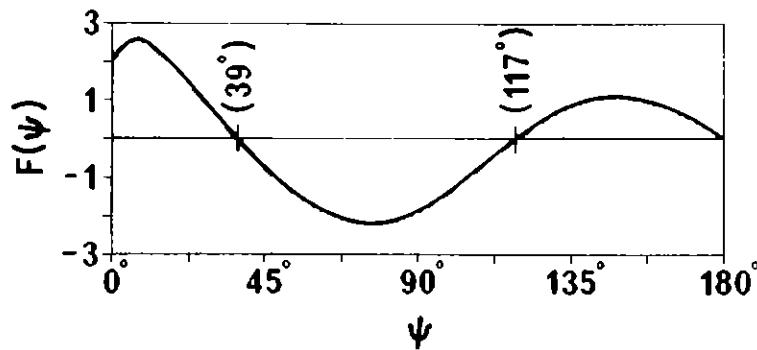


FIG. 22.17. Polar coordinates on the unit sphere.

FIG. 22.18.  $F$  function.

If  $f(\psi)$  can, in addition, be considered at least piece-wise constant, then the cumulative function  $\chi(\psi)$  of  $F(\psi)$  may be used to advantage. Denoting

$$\chi(\psi) = \int_0^\psi F(x) dx, \quad (22.74)$$

we again obtain an extremely well behaved function, and its graph is given in FIG. 19, according to LAMBERT AND DARLING [1936].

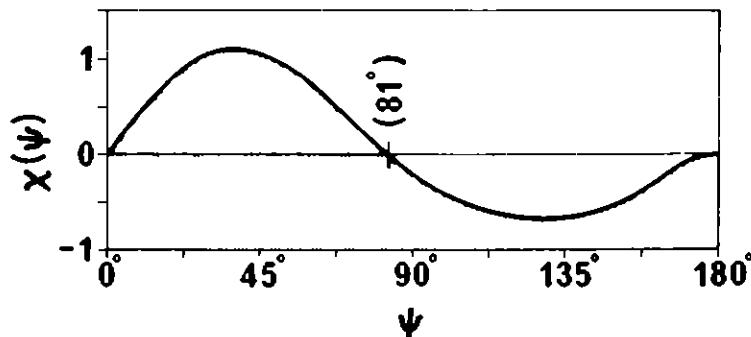
Similar treatment can be given to the Vening Meinesz kernel [SOLLINS, 1947], but this treatment will not be pursued here. Instead, it will be shown how the singularity is treated if geodetic rather than polar coordinates are employed, this time using the Vening Meinesz kernel as an example. In this case, the removal of the singularity calls for an entirely different approach.

For simplicity, let us deal with only the meridian component of the deflection of the vertical for which we can write (cf. (24))

$$\xi(\phi_A, \lambda_A) = \frac{1}{4\pi\gamma_0} \int_{\lambda=0}^{2\pi} \int_{\phi=-\pi}^{\pi} \Delta g(\phi, \lambda) \cos \alpha \frac{dS(\psi)}{d\psi} \cos \phi d\phi d\lambda. \quad (22.75)$$

Let us further assume that the point of interest ( $A$ ) is located in the block  $\mathcal{D}_1$  whose limits are  $\phi_1, \phi_2, \lambda_1, \lambda_2$ —see FIG. 20. This block is called the *innermost block* and requires a special treatment because of the singularity of the kernel:

$$K(\phi_A, \lambda_A, \phi, \lambda) = \frac{dS(\psi(\phi_A, \lambda_A, \phi, \lambda))}{d\psi} \cos \alpha(\phi_A, \lambda_A, \phi, \lambda) \cos \phi. \quad (22.76)$$

FIG. 22.19.  $\chi$  function.

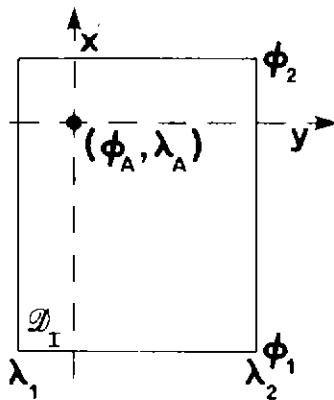


FIG. 22.20. Innermost block.

Hence the integral (75) is split into two integrals, one covering the innermost block  $\mathcal{D}_1$  and the other covering the rest of the earth's surface  $\mathcal{D}_O$ :

$$\xi = \xi_1 + \xi_O = \frac{1}{4\pi\gamma_0} \iint_{\mathcal{D}_1} K \Delta g d\phi d\lambda + \frac{1}{4\pi\gamma_0} \iint_{\mathcal{D}_O} K \Delta g d\phi d\lambda. \quad (22.77)$$

In the innermost block, Vening Meinesz's function can be developed into a power series in  $\psi$  to give

$$\frac{dS(\psi)}{d\psi} = -\frac{2}{\psi^2} - \frac{3}{\psi} - \frac{9}{2} + 8\psi + \dots \quad (22.78)$$

If the innermost block is no larger than, say,  $30' \times 30'$ , then  $\psi < 0.014$  radians and, to an accuracy better than 0.045%, it suffices to take only the first two terms in (78). Switching to Cartesian coordinates defined by (65), with their origin coinciding with  $A$ , one gets

$$\xi_1 = -\frac{1}{4\pi\gamma_0} \int_{y_1}^{y_2} \int_{x_1}^{x_2} \left( \Delta g \frac{x}{p^3} + \frac{3}{2} \Delta g \frac{x}{Rp^2} \right) dx dy, \quad (22.79)$$

where  $p = (x^2 + y^2)^{1/2}$ , so that  $\psi$  becomes  $p/R$ .

The gravity anomaly in  $\mathcal{D}_1$  may now be approximated by a surface:

$$\Delta g(x, y) = \Delta g_A + x \Delta g_x + y \Delta g_y + \dots \quad (22.80)$$

A good enough approximation, for a sufficiently small  $\mathcal{D}_1$ , can be obtained using just the first three terms, i.e., by regarding the anomaly as modellable by a plane. Clearly,  $\Delta g_A$  is the anomaly value at  $A$ , and  $\Delta g_x, \Delta g_y$  are the horizontal gradients of the anomaly in the meridian and prime vertical directions. Substitution for  $\Delta g$  from (80) back into (79) yields six additive terms which, upon integration, give

$$\begin{aligned} \xi_1 &\doteq -\frac{1}{4\pi\gamma_0} (\Delta g_A e_{11} + \Delta g_x e_{21} + \Delta g_y e_{31}) \\ &\quad - \frac{1}{4\pi\gamma_0 R} (\Delta g_A e_{12} + \Delta g_x e_{22} + \Delta g_y e_{32}) \\ &= -\frac{1}{4\pi\gamma_0} (\Delta g_A e_1 + \Delta g_x e_2 + \Delta g_y e_3). \end{aligned} \quad (22.81)$$

The coefficients  $e_1$ ,  $e_2$ , and  $e_3$  are all functions of  $R$  and the limits  $x_1, x_2, y_1, y_2$  [MERRY AND VANIČEK, 1974a]. A completely analogous equation can be derived for the prime vertical component  $\eta$ .

In practice, integration over the *outer blocks*  $\mathfrak{D}_O$  is replaced by summation of the contributions from individual cells  $\mathfrak{D}_i$  in the block. In this sense, the integration can be regarded as a summation over appropriately weighted mean anomalies, where the *weight kernel*  $w$  is given as

$$w(\phi_A, \lambda_A, \phi, \lambda) = \frac{dS(\psi)}{d\psi} \left\{ \begin{array}{l} \cos \alpha \\ \sin \alpha \end{array} \right\} \cos \phi. \quad (22.82)$$

The weight is evaluated for the centre of the corresponding cell  $\mathfrak{D}_i$ . The situation is slightly more complicated for cells closer to the computation point than about  $1.5^\circ$ . For these cells, the value of  $w$  pertaining to the centre of the block would not be accurate enough, and the weight must be calculated as an integral mean over the cell  $\mathfrak{D}_i$ , i.e.,

$$w(\phi_A, \lambda_A, \phi_i, \lambda_i) = \frac{1}{D_i} \iint_{\mathfrak{D}_i} \frac{dS(\psi)}{d\psi} \left\{ \begin{array}{l} \cos \alpha \\ \sin \alpha \end{array} \right\} \cos \phi d\mathfrak{D}, \quad (22.83)$$

where  $D_i$  is the area of the cell.

Sometimes the convolution integrals cannot be, or simply are not, evaluated through integration over the whole globe  $\mathbb{S}$ . Instead, the integration is carried out only over a circular cap  $\mathbb{S}_c$  of spherical radius  $\psi_c$ ; thus, for example, we have

$$\iint_{\mathbb{S}} \Delta g S(\psi) d\nu \doteq \iint_{\mathbb{S}_c} \Delta g S(\psi) d\nu. \quad (22.84)$$

This ‘truncated’ integration introduces, naturally, an error whose size depends on  $\psi_c$ . To reduce this error as much as possible, one can appropriately modify the integration kernel. There are several ways of modifying the kernel, the best known being through the introduction of *truncation coefficients*  $Q_n$  [MOLODENSKIJ ET AL., 1960]. Stokes’s function modified in this way reads (cf. (15))

$$\tilde{S}(\psi) = \sum_{n=0}^{\infty} \frac{2n+1}{2} Q_n(\psi_c) P_n(\cos \psi), \quad (22.85)$$

where

$$Q_n(\psi_c) = \int_{-1}^{\cos \psi_c} S(\xi) P_n(\xi) d\xi. \quad (22.86)$$

Other alternatives may be found in, e.g., JEKELI [1980].

As the reader should appreciate by now, all the convolution integrals derived in this chapter give us only point values of the gravity field parameters being sought. In other words, each integration over the whole earth’s surface gives us only one value of the parameter. The question one may ask is: Is there a way of using these formulae to simultaneously evaluate the desired quantity for a whole region? Not directly. If a global knowledge of the parameter is needed, then it becomes

preferable to evaluate the potential coefficients from the global anomalies (cf. (12)) and generate the desired quantity from a spherical harmonic series. Sometimes, however, we are interested in computing the desired parameter only for a small region of the earth's surface and, if a reasonably detailed account of the parameter is desired, an inappropriately large number of terms in the harmonic series would have to be taken into consideration. If such *regional determination of a gravity field parameter* is needed, one might want to seek the quantity in question, be it geoidal height, deflection component, etc., in the form of a generalized two-dimensional polynomial (surface) valid only for that region.

This approach clearly parallels the prediction of anomalies by regression discussed in §22.3. To develop the technique further, let us denote the parameter being sought by  $P$  and model it in the following form:

$$P(\phi_A, \lambda_A) \doteq \tilde{\Phi}^T(\phi_A, \lambda_A) \mathbf{c}, \quad (22.87)$$

where  $\tilde{\Phi}^T(\phi_A, \lambda_A)$  is one column of Vandermonde's matrix belonging to the  $n$  selected base functions (algebraic or other), and  $\mathbf{c}$  is an  $n$  vector of coefficients to be determined. On the other hand, for the same parameter we can write the following integral:

$$P(\phi_A, \lambda_A) = \iint_S f(\phi, \lambda) K(\phi_A, \lambda_A, \phi, \lambda) d\nu, \quad (22.88)$$

where  $K$  is the appropriate kernel, and  $f$  is the known (observed) function, typically the gravity anomaly of one kind or another.

Now, the right-hand sides of both equations can be approximately equated on a selected mesh of  $m$  points ( $\mathfrak{T} \equiv \{\phi_i, \lambda_i; i \in I\}$ ), covering the region of interest, and we obtain

$$\tilde{\Phi}^T(\phi_i, \lambda_i) \mathbf{c} \doteq \iint_S f(\phi, \lambda) K(\phi_i, \lambda_i, \phi, \lambda) d\nu, \quad i = 1, \dots, m. \quad (22.89)$$

These  $m$  equations can be rewritten in a matrix form as follows:

$$\tilde{\Phi}^T(\mathfrak{T}) \mathbf{c} \doteq \iint_S f(\phi, \lambda) \mathbf{K} d\nu, \quad (22.90)$$

where  $\tilde{\Phi}^T(\mathfrak{T})$  is now the full Vandermonde matrix, and  $\mathbf{K}$  is a vector-kernel with  $m$  components. Provided  $m > n$ , we have a linear problem that can be solved for  $\mathbf{c}$  using the least-squares method. Denoting the covariance matrix of  $P$  as evaluated on the mesh-points by  $\mathbf{C}_p$ , we get

$$[\tilde{\Phi}^T(\mathfrak{T}) \mathbf{C}_p^{-1} \tilde{\Phi}^T(\mathfrak{T})] \hat{\mathbf{c}} = \tilde{\Phi}^T(\mathfrak{T}) \mathbf{C}_p^{-1} \iint_S f \mathbf{K} d\nu. \quad (22.91)$$

Inverting the matrix  $\tilde{\Phi}^T(\mathfrak{T}) \mathbf{C}_p^{-1} \tilde{\Phi}^T(\mathfrak{T})$  and interchanging the matrix multiplication with the integration, we finally obtain the estimated coefficients as

$$\begin{aligned} \hat{\mathbf{c}} &= \iint_S f(\phi, \lambda) [\tilde{\Phi}^T(\mathfrak{T}) \mathbf{C}_p^{-1} \tilde{\Phi}^T(\mathfrak{T})]^{-1} \tilde{\Phi}^T(\mathfrak{T}) \mathbf{C}_p^{-1} \mathbf{K} d\nu \\ &= \iint_S f(\phi, \lambda) \tilde{\mathbf{K}} d\nu, \end{aligned} \quad (22.92)$$

where  $\tilde{\mathbf{K}}$  is a new anisotropic vector-kernel. Hence the estimated coefficients  $\hat{c}_i$  can be evaluated directly from their own global surface convolution integrals much the same way as, for example, the potential coefficients. Note that although the density and configuration of the mesh do not appear explicitly in the kernel, they do influence its form as does the selection of the base functions and the covariance matrix  $\mathbf{C}_p$ .

Clearly, the problem of regional prediction of the desired parameter  $P$  on the basis of observed gravity can also be handled through collocation or, better yet, through a combination of regression and collocation. The covariance functions of  $N, \xi, \eta$  can be obtained from the covariance function of  $\Delta g$  (see §22.3) analytically. The reader can find the pertinent equations in, e.g., LACHAPELLE [1978] and TSCHERNING AND FORSBERG [1978]. Another technique using the combination of spherical harmonics and surface integrals will be shown in §24.4.

The last point to discuss in this chapter is the evaluation of errors in the determined field parameters. To track down these errors, let us write the value of the sought parameter  $P$  at point  $A$  in the following form:

$$P(\phi_A, \lambda_A) = \sum_i f(\phi_i, \lambda_i) w(\phi_A, \lambda_A, \phi_i, \lambda_i), \quad (22.93)$$

where  $f$  is the observed function, and  $w$  is the weight kernel. This equation can be rewritten in matrix form as

$$P = \mathbf{w}^T \mathbf{f}. \quad (22.94)$$

Denoting now the covariance matrix of the observed function by  $\mathbf{C}_f$  and applying the covariance law, we obtain for the standard error in  $P$

$$\sigma_p = (\mathbf{w}^T \mathbf{C}_f \mathbf{w})^{1/2}. \quad (22.95)$$

The covariance matrix  $\mathbf{C}_f$  can be assembled from the variances of the individual values of  $f$  and their covariance function (e.g., (68) or FIG. 16) following the methodology outlined in Chapter 10. Clearly, if the contribution of the observed data from the immediate neighbourhood of  $A$  was calculated separately, then the error contribution of the immediate neighbourhood of  $A$  also has to be evaluated separately by propagating the errors in  $f$  through the particular formula used. Determination of the  $\mathbf{C}_p$  of a whole string of parameter values calculated either from the point determination formulae or from (87) or (92) is left to the reader.

Practical results indicate, as one may expect, that the accuracy of the calculated parameters varies widely with location. In regions where gravity coverage is dense and homogeneous around the computation point, the resulting geoidal height or height anomaly can be determined accurately to a few metres. Correspondingly, the deflection components would be good to one or two seconds of arc. On the other hand, in the vicinity of larger blank areas or areas with unreliably predicted anomalies, the accuracy may be worse by as much as one order of magnitude. Also, the size of cells used in the computations affects the accuracy—the larger the cells, the worse the errors. For more in-depth treatment and generalized error formulae, the reader is referred to publications such as KAULA [1959], GROten AND MORITZ [1964], and MORITZ [1979].

## CHAPTER 23

### DETERMINATION OF THE GRAVITY FIELD FROM OBSERVATIONS TO SATELLITES

The determination of the gravity field parameters from observations to satellites requires an understanding of how satellites move in a gravitational field. To study this motion, it is necessary to make use of yet another part of physics—dynamics (see §2.3); throughout this chapter, we shall assume that the basics of dynamics are known to the reader. To simplify the mathematical coding, we shall use the dot above a symbol to denote the rate of change with time (velocity) of the quantity in question. A subscript to the  $\nabla$  operator will signify the coordinate system in which the operator is to be evaluated (see §3.2).

In the first section, we describe the fundamentals of satellite dynamics with the emphasis on high orbiting satellites. As well, the concept of the determination of the mass of the earth is also shown. The second section contains the basics of low orbiting satellite dynamics and satellite orbit prediction. The inverse problem of satellite dynamics, i.e., the determination of gravity field parameters, is dealt with in the last two sections. Section three introduces the basic mathematical tool—the perturbation theory. Particular attention is paid to secular perturbations and to the evaluation of the flattening of the earth. The last section then treats the concepts needed for the determination of the remaining parameters.

#### 23.1. Satellites and the gravitational field

Both natural and artificial earth satellites can be conveniently divided into two groups according to their altitude. When their altitude is much higher than the value of the earth's radius, we speak of *high orbiting satellites*. The rest are referred to as low orbiting or *close satellites*. An extreme case of high orbiting satellites are space probes designed to escape from the earth's gravitational field altogether.

The motion of a satellite is governed by its own kinetic energy and the forces that act on it. In the vicinity of the earth, the predominant force acting on the satellite is the earth's gravitational pull, and the satellite responds by following an orbit that reflects the shape of the gravitational field. In very high altitudes, the gravitational field degenerates into an approximately radial field. This fact can be readily seen from, e.g., (20.61), where the second (elliptical) term diminishes much more rapidly

with growing distance  $r$  than the first (radial) term. In lower altitudes, the field is rather irregular with the elliptical term representing the most significant departure from radiality—see FIG. 1. In fact, the lower the altitude, the more irregular the field.

To adequately describe the orbit (motion)—cf. §15.3—of a high orbiting satellite, or a *space probe*, it is possible to use the classical Keplerian theory mentioned in §5.1. The theory ascertains that the satellite orbit in a radial field is planar so that five out of the six Keplerian orbital elements (see §15.3)  $\mathbf{k} = \{a_0, e, i, \mu, \omega, \Omega\}$  remain constant, and only one, the mean anomaly  $\mu$ , changes with time (linearly). This is not, however, the case with close satellites which, as we shall see in the next section, have to be treated differently.

The motion of a satellite in a radial field can be fully described by the *angular-momentum integral* [KOVALEVSKY, 1967]. This integral relates the kinetic energy  $T$  (not to be confused with the disturbing potential) of the satellite with the radial field characteristics:

$$T(r) = \frac{\dot{r}^2}{2} = \frac{GM}{2} \left( \frac{2}{r} - \frac{1}{a_0} \right), \quad (23.1)$$

where  $\dot{r}$  is the rate of change of  $r$  with time, and  $a_0$  is the major semi-axis of the satellite orbital ellipse (see FIG. 2). This formula is valid not only for radial but any other field and is extensively used in *celestial mechanics*. Since it directly relates the geometry of the orbit and the satellite velocity with the mass of the earth, it can be used in evaluating  $M$ —or, more precisely,  $GM$ —from observations to high orbiting satellites. The most accurate determination of  $GM$  based on this approach uses

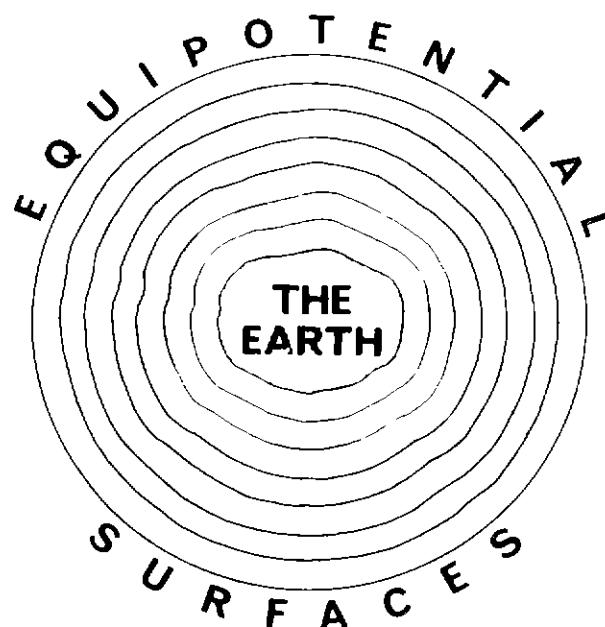


FIG. 23.1. Departures of the actual gravity field from radicity.

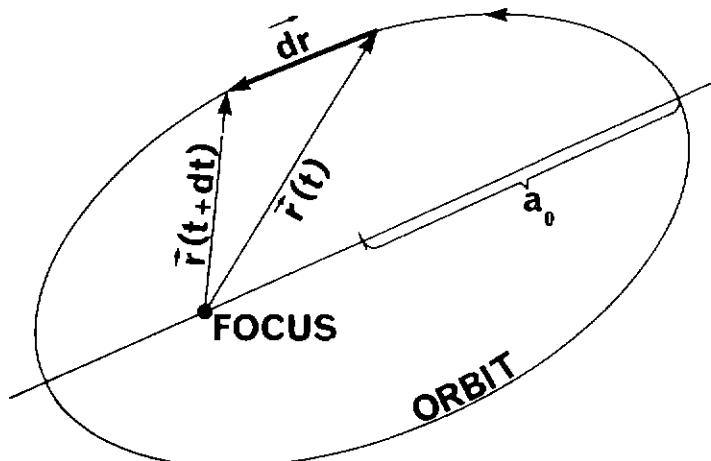


FIG. 23.2. Keplerian orbit.

space-probe tracking [ESPOSITO AND NG, 1976]. It has to be noted that the mass of the earth determined in this way includes the mass of the atmosphere as well (cf. §9.4).

Since the deviations from radiality of the actual field are comparatively small, even the deviations of the actual orbit from the Keplerian orbit are relatively small. Thus, when dealing with a close satellite and its *non-Keplerian orbit*, it is convenient to treat the orbit as a perturbed Keplerian orbit. To do this, we usually write the potential of the field as

$$W(\vec{r}) = \frac{GM}{r} + P(\vec{r}), \quad (23.2)$$

where  $P$  is the part of the actual potential describing the departures from radiality. It is called the *perturbing potential*, and it represents the non-radial contribution of the earth's gravitational field given, e.g., by the higher order harmonics in (20.58). Although the perturbing potential is different from the disturbing potential (cf. 20.90), they are closely related, as we shall see in §23.4.

The non-gravitational part of the orbital perturbation consists of perturbations due to air drag, electromagnetic forces, solar radiation pressure, tidal forces and relativistic effects.

(a) *Air drag* is the force caused by air friction; it acts against the motion of the satellite, and it is proportional to the satellite's velocity. It is evidently more important for closer satellites, and for these it is the most significant of all the non-gravitational forces. Air drag is difficult to model, and this is why so-called *drag-free satellites* are now being contemplated. A drag-free satellite is really an inner satellite suspended within an outer shell. When the outer shell is forced to follow the same trajectory as the inner body, the effect of the drag is clearly eliminated.

(b) The *electromagnetic force perturbation* is of a similar nature; it is caused by the interaction of the satellite's electrical charge, acquired by passing through the ionosphere, with the earth's magnetic field.

(c) *Solar radiation pressure effect* is caused by a force acting in the direction from the sun and has a magnitude proportional to the ratio of the cross section area to the density of the satellite. It is thus particularly significant for light and large satellites at high altitudes.

(d) *Tidal perturbations* are caused by forces which have already been treated in §8.1 and are relatively easy to account for, as we shall see in Chapter 25. More complicated is the evaluation of the indirect effect of oceanic tide [LAMBECK ET AL., 1974].

(e) *Relativistic effects* arise from the fact that satellite velocities, typically about  $3 \times 10^{-4}$  times the velocity of light, are relatively high and thus should be described using relativistic dynamics. Because Newtonian dynamics is normally used, corrections have to be formulated for this shortcoming. A more detailed treatment of non-gravitational effects can be found in KING-HELE [1967] and MORANDO [1970].

Of course, some of the non-gravitational forces act even on very high orbiting satellites. Thus these orbits must also be corrected before they can be treated as Keplerian. It should be noted that the earth's centrifugal force does not affect the satellite motion because the satellites are not rigidly attached to the earth; they orbit freely above the spinning earth. Finally, it should be stated that often it is expedient to consider the gravitational perturbing potential  $P$  as being composed of three parts: one  $P_E$  due to the ellipticity of the earth, another  $P_Z$  due to the *zonal irregularities* (cf. §20.2), and the last  $P_T$  due to the *tesseral irregularities*. The reason for this classification will become obvious in the next section.

## 23.2. Prediction of orbits

It is intuitively clear that it would be quite difficult to formulate the equations of motion (i.e., three ordinary differential equations of second order relating the satellite acceleration components with components of the forces acting on it) of a close satellite using the standard approach. It is much simpler to use the tools of *analytical mechanics*, such as generalized coordinates, the Hamiltonian function, and canonical equations of motion [SYMON, 1971].

Since the motion of a satellite in a given potential field is fully described by three ordinary differential equations of second order, three *generalized coordinates*—e.g., any curvilinear coordinates  $\mathbf{q} \equiv \{q_1, q_2, q_3\}$  (see §3.3)—can be selected to describe the motion. The corresponding *generalized momenta*  $\mathbf{p}$  are then given as

$$\mathbf{p} = \nabla_{\dot{\mathbf{q}}} (T + W), \quad (23.3)$$

where  $T$  is the kinetic energy of the satellite, and  $W$  is the potential of the field. It can be shown that if  $W$  does not depend explicitly on time in the  $\mathbf{q}$  coordinate

system, then the following *canonical equations of motion* are valid:

$$\dot{\mathbf{p}} = -\nabla_{\mathbf{q}} H, \quad \dot{\mathbf{q}} = \nabla_{\mathbf{p}} H, \quad (23.4)$$

where

$$H = T - W \quad (23.5)$$

is called the *Hamiltonian function* of the satellite. Equations (4) are six ordinary differential equations of first order in the  $\{\mathbf{p} \mid \mathbf{q}\}$  coordinate system, mathematically equivalent to the standard three ordinary differential equations of second order required to describe the motion.

It has been shown by Delaunay [KOVALEVSKY, 1967] that the last three orbital elements,  $\mathbf{q} \equiv \{\mu, \omega, \Omega\}$ , can be used as generalized coordinates. The generalized momenta corresponding to these coordinates are

$$\mathbf{p} = \sqrt{GMa_0} \{1, \nu, \nu \cos i\}, \quad (23.6)$$

where  $\nu^2 = 1 - e^2$ . The generalized momenta are clearly functions of only the remaining three orbital parameters  $a_0, e, i$ . Note that this particular generalized coordinate system is not suitable for a *circular orbit*, defined by  $e=0$ , or an *equatorial orbit*, defined by  $i=0$  for which  $p_2 = p_1$  and  $p_2 = p_3$  respectively, and these momenta cannot be used as coordinates in canonical equations of motion. For these special orbits, we can use, for instance, the following modified coordinate system [KAULA, 1962]:

$$\mathbf{q}' \equiv \{q_1 + q_2 + q_3, q_2 + q_3, q_3\} \quad (23.7)$$

that yields

$$\mathbf{p}' = \{p_1, p_2 - p_1, p_3 - p_2\}. \quad (23.8)$$

In the present development however, only original *Delauney's coordinates* will be dealt with.

Disregarding the non-gravitational effects for the time being, the Hamiltonian function becomes

$$H = \frac{GM}{2} \left( \frac{2}{r} - \frac{1}{a_0} \right) - \left( \frac{GM}{r} + P \right), \quad (23.9)$$

where the kinetic energy is given by the angular-momentum integral (1), and the gravitational potential is taken from (2). Expressing  $H$  in the  $\{\mathbf{p} \mid \mathbf{q}\}$  coordinate system, we get

$$H(\mathbf{p}, \mathbf{q}) = -\frac{1}{2} \left( \frac{GM}{p_1} \right)^2 - P(\mathbf{p}, \mathbf{q}). \quad (23.10)$$

The canonical equations of motion can now be easily derived; they read

$$\boxed{\dot{\mathbf{p}} = \nabla_{\mathbf{q}} P}, \quad (23.11)$$

$$\dot{q}_1 = \sqrt{\frac{GM}{a_0^3}} - \frac{\partial P}{\partial p_1}, \quad \dot{q}_2 = -\frac{\partial P}{\partial p_2}, \quad \dot{q}_3 = -\frac{\partial P}{\partial p_3}. \quad (23.12)$$

To make the two sets of equations ((11) and (12)) compatible, let us denote

$$\delta\mu = \mu - \sqrt{GM/a_0^3}, \quad (23.13)$$

where  $\sqrt{GM/a_0^3}$  can be shown to be the mean anomaly of a Keplerian motion with orbital parameters  $a_0, e, i, \omega, \Omega$ . Denoting the triplet  $\{\delta\mu, \omega, \Omega\}$  by  $\tilde{\mathbf{q}}$ , we can rewrite the second set of canonical equations (12) as

$$\boxed{\dot{\tilde{\mathbf{q}}} = -\nabla_{\mathbf{p}} P}. \quad (23.14)$$

Equations (11) and (14) are the basic equations of motion used in close satellite dynamics. Note that for a radial field ( $P=0$ ), zero time changes are obtained in either  $\tilde{\mathbf{q}}$  or  $\mathbf{p}$ , and the motion becomes Keplerian as expected.

The two derived systems of equations are still awkward to handle because of the presence of generalized momenta  $\mathbf{p}$ . For this reason, they are usually reformulated in terms of orbital parameters alone. Let us denote  $\{a_0, e, i\}$  by  $s$  and  $\tilde{\mathbf{k}} \equiv \{s \mid \tilde{\mathbf{q}}\} \equiv \{a_0, e, i, \delta\mu, \omega, \Omega\}$ . Then we can write

$$\frac{\partial P}{\partial p_j} = \sum_{i=1}^3 \frac{\partial s_i}{\partial p_j} \frac{\partial P}{\partial s_i}, \quad j=1,2,3, \quad (23.15)$$

or simply

$$\nabla_{\mathbf{p}} P = \frac{\partial s}{\partial \mathbf{p}} \nabla_s P, \quad (23.16)$$

where  $\partial s / \partial \mathbf{p}$  is the Jacobian of transformation (cf. §3.1) between  $s$  and  $\mathbf{p}$  coordinate systems. On the other hand,

$$\dot{p}_j = \frac{d p_j}{d \tau} = \sum_{i=1}^3 \frac{\partial p_j}{\partial s_i} \frac{ds_i}{d \tau} = \sum_{i=1}^3 \frac{\partial p_j}{\partial s_i} \dot{s}_i, \quad j=1,2,3, \quad (23.17)$$

or

$$\dot{\mathbf{p}} = \frac{\partial \mathbf{p}}{\partial s} \dot{\mathbf{s}} = \left( \frac{\partial s}{\partial \mathbf{p}} \right)^T \dot{\mathbf{s}}. \quad (23.18)$$

Hence

$$\dot{\tilde{\mathbf{k}}} = \begin{bmatrix} \dot{\mathbf{s}} \\ \dot{\tilde{\mathbf{q}}} \end{bmatrix} = \begin{bmatrix} \frac{\partial s}{\partial \mathbf{p}} & \mathbf{0} \\ -\frac{\partial \tilde{\mathbf{q}}}{\partial s} & -\frac{\partial s}{\partial \mathbf{p}} \end{bmatrix} \begin{bmatrix} \nabla_{\mathbf{q}} P \\ \nabla_s P \end{bmatrix} = \mathbf{B} \nabla_k P. \quad (23.19)$$

It is left to the reader to show that

$$\frac{\partial \mathbf{s}}{\partial \mathbf{p}} = \frac{1}{\sqrt{GMa_0}} \begin{bmatrix} 2a_0 & \frac{\nu^2}{e} & 0 \\ 0 & \frac{-\nu}{e} & \frac{\cot i}{\nu} \\ 0 & 0 & -\frac{1}{\nu \sin i} \end{bmatrix}. \quad (23.20)$$

The reformulated equations of motion then read in full

$$\dot{a}_0 = (GMa_0)^{-1/2} 2a_0 \frac{\partial P}{\partial \mu},$$

$$\dot{e} = (GMa_0)^{-1/2} \left( \frac{\nu^2}{e} \frac{\partial P}{\partial \mu} - \frac{\nu}{e} \frac{\partial P}{\partial \omega} \right),$$

$$\dot{i} = (GMa_0)^{-1/2} \left( \frac{\cot i}{\nu} \frac{\partial P}{\partial \omega} - \frac{1}{\nu \sin i} \frac{\partial P}{\partial \Omega} \right),$$

$$\dot{\mu} = (GMa_0)^{-1/2} \left( -2a_0 \frac{\partial P}{\partial a_0} - \frac{\nu^2}{e} \frac{\partial P}{\partial e} \right),$$

$$\dot{\omega} = (GMa_0)^{-1/2} \left( \frac{\nu}{e} \frac{\partial P}{\partial e} - \frac{\cot i}{\nu} \frac{\partial P}{\partial i} \right),$$

$$\dot{\Omega} = (GMa_0)^{-1/2} \frac{1}{\nu \sin i} \frac{\partial P}{\partial i}.$$

(23.21)

These equations can also be regarded as equations for the time rates or velocities of the Keplerian orbital parameters.

For studying the effect of non-gravitational *perturbing forces*, it is handy to express the orbital element velocities as direct functions of these forces  $\mathbf{F}$ . To develop this relation, let us first define a Cartesian orbital coordinate system  $\xi \equiv \{\rho, \kappa, \eta\}$  that moves with the satellite (see FIG. 3). Let the  $\rho$ -axis point toward the earth's centre of mass, the  $\kappa$ -axis point in the direction of motion in the instantaneous orbital plane, and the  $\eta$ -axis complete the right-handed system. Note that for small  $e$ , the  $\kappa$  direction practically coincides with the tangent to the orbit.

Now the following transformation equations for gradients can be written:

$$\nabla_s = \frac{\partial \xi}{\partial s} \nabla_\xi, \quad \nabla_q = \frac{\partial \bar{\xi}}{\partial q} \nabla_\xi. \quad (23.22)$$

Combining these equations with (19), we arrive at the following equations:

$$\dot{\mathbf{s}} = - \frac{\partial \mathbf{s}}{\partial \mathbf{p}} \frac{\partial \xi}{\partial \mathbf{q}} \mathbf{F}, \quad (23.23)$$

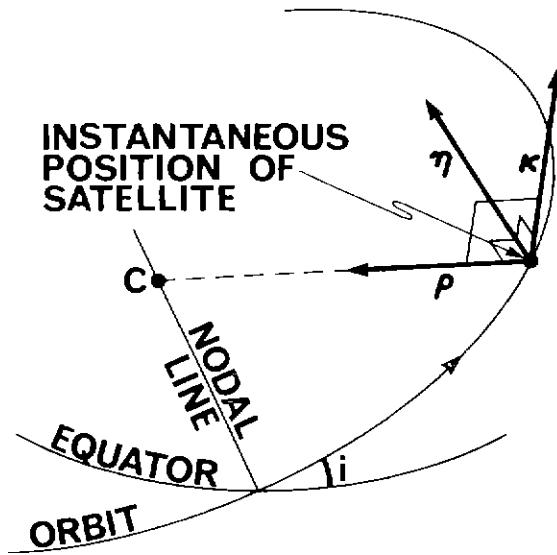


FIG. 23.3. Instantaneous orbital Cartesian system of coordinates.

$$\dot{\vec{q}} = \frac{\partial s}{\partial p} \frac{\partial \xi}{\partial s} \mathbf{F}, \quad (23.24)$$

linking the perturbing forces with the orbital element velocities. Finally we get

$$\dot{\vec{k}} = \begin{bmatrix} -\frac{\partial s}{\partial p} & \frac{\partial \xi}{\partial s} \\ -\frac{\partial p}{\partial s} & \frac{\partial \xi}{\partial p} \\ \frac{\partial s}{\partial \xi} & \frac{\partial p}{\partial s} \end{bmatrix} \mathbf{F} = \mathbf{DF}. \quad (23.25)$$

Evaluating the appropriate Jacobians, we obtain [TUCKER ET AL., 1970]

$$\begin{aligned} \dot{a}_0 &= Q \left( \frac{2a_0^2 e}{r} \sin f F_p + \frac{2a_0^3 v^2}{r^2} F_\kappa \right), \\ \dot{e} &= Q \left[ \frac{a_0 v^2}{r} \sin f F_p + \left( e + \left( 1 + \frac{a_0 v^2}{r} \right) \cos f \right) F_\kappa \right], \\ \dot{i} &= Q \cos(\varpi + f) F_\eta, \\ \dot{\delta\mu} &= Q v \left[ \left( 2 - \frac{a_0 v^2}{re} \cos f \right) F_p - \left( 1 + \frac{a_0 v^2}{r} \right) \frac{\sin f}{e} F_\kappa \right], \\ \dot{\varpi} &= Q \left[ -\frac{a_0 v^2}{re} \cos f F_p + \left( 1 + \frac{a_0 v^2}{r} \right) \frac{\sin f}{e} F_\kappa - \frac{\sin(\varpi + f)}{\tan i} F_\eta \right], \\ \dot{\omega} &= Q \frac{\sin(\varpi + f)}{\sin i} F_\eta, \end{aligned} \quad (23.26)$$

where  $Q = (r/v)(GMa_0)^{-1/2}$ , and  $f$  is the true anomaly defined in §15.3. An alternative derivation of these equations, based on a simple vector description of orbital dynamics, can be found in POLLARD [1976].

If the perturbing potential  $P$  is known, and the non-gravitational effects are disregarded, it is a simple matter to *predict the orbit*, i.e., to predict the six orbital elements of a satellite for any instant  $\tau$ . Knowing the ‘position’  $\tilde{\mathbf{k}}$  of the satellite at an initial instant  $\tau_0$ , the position at  $\tau$  is

$$\tilde{\mathbf{k}}(\tau) = \tilde{\mathbf{k}}(\tau_0) + \int_{\tau_0}^{\tau} \dot{\tilde{\mathbf{k}}} d\tau = \tilde{\mathbf{k}}(\tau_0) + \int_{\tau_0}^{\tau} \mathbf{B}(\tau) \nabla_k P d\tau. \quad (23.27)$$

The integration can be carried out numerically or analytically (as shall be done here). If the perturbing potential is adequately known, the prediction can extend farther—up to several revolutions of the satellite—and we speak of a *long-arc prediction*.

*Short-arc prediction* can be done with or without an explicit knowledge of  $P$ . If  $P$  is inadequately known, the individual orbital-element variations can be extrapolated using, for instance, a generalized polynomial  $\tilde{\Phi}_i^T(\tau)\lambda$  whose coefficients  $\lambda$  are evaluated from the known (observed) part of the orbit (cf. §14.2). The predicted values for  $\tau$  are then obtained from

$$\tilde{k}_i(\tau) = \tilde{k}_i(\tau_0) + \tilde{\Phi}_i^T(\tau - \tau_0)\lambda, \quad i = 1, \dots, 6. \quad (23.28)$$

For an example, see VEIS AND MOORE [1960].

Orbital parameters are by no means the only usable coordinates for orbital prediction. Various Cartesian coordinate systems may be and are used for this task. It must also be pointed out that the  $P_T$  part of the perturbing potential  $P$  is not explicitly independent of time as required by the Hamiltonian theory; as such, its contribution to  $P$  must be appropriately corrected. These questions are, however, considered beyond the scope of this outline and the interested reader is referred to, e.g., KAULA [1966].

### 23.3. Analysis of orbital perturbations

Let us suppose, for the time being, that all the non-gravitational effects have been taken care of as corrections to the observed orbit. Then the departure of this corrected observed orbit from a Keplerian orbit would be due only to the irregularities of the earth’s gravitational field. The integrated departures, known as *orbital perturbations*, can be obtained directly from (27) in terms of orbital elements:

$$\delta\tilde{\mathbf{k}}(\tau) = \tilde{\mathbf{k}}(\tau) - \tilde{\mathbf{k}}(\tau_0) = \int_{\tau_0}^{\tau} \mathbf{B}(\tau) \nabla_k P d\tau. \quad (23.29)$$

This equation gives us the mathematical model linking the observed perturbations  $\delta\tilde{\mathbf{k}}$  with the gravitational perturbing potential  $P$ . In principle,  $P$  can be evaluated from

(29). Practically, however, this proposition—really a problem inverse to orbital prediction—presents some formidable difficulties.

To begin with,  $P$  must be expressed in orbital parameters  $\mathbf{k}$  to enable us to evaluate the gradient  $\nabla_{\mathbf{k}} P$ . In geocentric coordinates  $\bar{r} \equiv \{r, \phi, \lambda\}$ , the perturbing potential reads (cf. (20.58))

$$P(\bar{r}) = -\frac{GM}{r} \sum_{n=2}^{\infty} \left(\frac{a}{r}\right)^n \sum_{m=0}^n (J_{nm} \cos m\lambda + K_{nm} \sin m\lambda) P_{nm}(\sin \phi). \quad (23.30)$$

Transformation of this expression into orbital parameters is very laborious, and we shall content ourselves here with quoting just the final result [KAULA, 1966; CAPUTO, 1967]:

$$P(\mathbf{k}) = \frac{GM}{a_0} \sum_{n=2}^{\infty} \left(\frac{a}{a_0}\right)^n \sum_{m,p=0}^n F_{nmp}(i) \sum_{q=-\infty}^{\infty} G_{mpq}(e) S_{nmpq}(\mu, \varpi, \omega, \theta).$$

$$(23.31)$$

In both the above equations,  $a$  once more is the major semi-axis of the geocentric reference ellipsoid;  $F_{nmp}(i)$  and  $G_{mpq}(e)$  are complicated functions of  $i$  and  $e$ . The function  $S_{nmpq}$  is of the following form:

$$S_{nmpq} = \begin{cases} -J_{nm} \cos \psi - K_{nm} \sin \psi, & n-m \text{ even}, \\ -J_{nm} \sin \psi + K_{nm} \cos \psi, & n-m \text{ odd}, \end{cases} \quad (23.32)$$

where

$$\psi = (n-2p)\varpi + (n-2p+q)\mu + m(\omega - \theta), \quad (23.33)$$

with  $\theta$  denoting the GAST (see §15.1) which describes the earth's spin rate underneath the orbit. It is illustrative to show the special forms (32) takes for  $P_E$  and  $P_Z$ . As the reader can easily verify,

$$P_E = -\frac{GMa^2}{a_0^3} J_2 \sum_{p=0}^2 \sum_{q=-\infty}^{\infty} F_{20p}(i) G_{2pq}(e) \cos[(2-2p)\varpi + (2-2p+q)\mu], \quad (23.34)$$

and

$$P_Z = -\frac{GM}{a_0} \sum_{n=3}^{\infty} \left(\frac{a}{a_0}\right)^n J_n \sum_{p=0}^n \sum_{q=-\infty}^{\infty} F_{n0p}(i) G_{npq}(e) \times \begin{cases} \cos[(2-2p)\varpi + (n-2p+q)\mu] \\ \sin[(2-2p)\varpi + (n-2p+q)\mu] \end{cases}, \quad \begin{cases} n \text{ even}, \\ n \text{ odd}. \end{cases} \quad (23.35)$$

Clearly,  $P$  is a linear function of potential coefficients  $J_{nm}$  and  $K_{nm}$ . Consequently, its gradient, taken with respect to  $\mathbf{k}$ , as well as the perturbations must be linear functions of these coefficients. Therefore, every observed perturbation of an orbital element gives us one linear observation equation for determining the potential

coefficients. Note that  $P$  could be alternatively expressed in, say, the CT system, and (29) could be reformulated to hold in the same system. In this approach, however, the advantage of using a Keplerian orbit for the 0th approximation to the real orbit would be lost.

Because the  $J_2$  coefficient is so much larger than the rest of the coefficients (cf. §20.2), a simultaneous solution for all of the coefficients would not be numerically very stable. For this reason,  $J_2$  is normally evaluated first, independent of the rest. When solving for  $J_2$ , we can see (34) that  $P_E$  (34) does not depend on  $\omega$ . It can also be shown that dependence of  $P_E$  on  $\mu$  is very weak so that, in the first approximation, it can be neglected [KAULA, 1966]. This is equivalent to saying that the coefficient  $(2 - 2p + q)$  by  $\mu$  in (34) goes to zero. This, in turn, indicates that only the following three combinations of  $p$  and  $q$  have to be considered: 0, -2; 1, 0; 2, 2. For the first and third combinations, however, we find that  $G_{20-2}(e) = G_{222}(e) = 0$ , which leaves only the second combination for consideration. For  $p = 1, q = 0$  we find

$$G_{210}(e) = \nu^{-3}, \quad F_{201}(i) = \frac{3}{4} \sin^2 i - \frac{1}{2}, \quad (23.36)$$

and the first approximation of  $P_E$  becomes

$$P_E \doteq \frac{GMa^2}{\nu^3 a_0^3} \left( \frac{3}{4} \sin^2 i - \frac{1}{2} \right) J_2. \quad (23.37)$$

From this equation it is clear that

$$\nabla_q P_E \doteq \mathbf{0}, \quad (23.38)$$

because, in the first approximation,  $P_E$  does not depend on either  $\mu$  or  $\omega$  or  $\omega$ . According to (19), a further consequence is that

$$\dot{s} \doteq \mathbf{0}, \quad (23.39)$$

and we have thus discovered that the first three orbital parameters  $a_0, e, i$  are not (in the first approximation) perturbed by the elliptical potential  $P_E$ . Since the hypermatrix  $B$  is a function of only these three parameters (cf. (21)),  $B$  does not change with time either if  $P_E$  is considered.

On the other hand, the gradient  $\nabla_s P_E$  becomes

$$\nabla_s P_E \doteq 3 \frac{GMa^2}{\nu^3 a_0^3} J_2 \left\{ \frac{\frac{3}{4} \sin^2 i - \frac{1}{2}}{a_0}, -e \frac{\frac{1}{4} \sin^2 i - \frac{1}{2}}{\nu^2}, -\frac{1}{4} \sin 2i \right\}. \quad (23.40)$$

Substitution into (21) and integration with respect to time finally yield

$$\begin{aligned} \delta\delta\mu(\tau) &\doteq Q \left( \frac{3}{4} \sin^2 i - \frac{1}{2} \right) J_2 \tau, \\ \delta\omega(\tau) &\doteq \frac{Q}{4\nu} (1 - 5 \cos^2 i) J_2 \tau, \\ \delta\omega(\tau) &\doteq \frac{Q}{2\nu} \cos i J_2 \tau, \end{aligned} \quad (23.41)$$

where  $Q = -(3/\nu^3)\sqrt{GM/a_0^3}(a/a_0)^2$ . These quantities are known as *secular perturbations* and are used for an approximate evaluation of  $J_2$ . The error in the first approximation to  $J_2$  is of the same order as the values of the rest of the potential coefficients; thus the correction to approximate  $J_2$  is solved for together with them.

Once the effect of the earth's ellipticity on the orbit has been estimated and corrected for, using the approximate value of  $J_2$ , the remaining potential coefficients can be evaluated. In showing how this is done, again we shall develop only the first approximation to the perturbation equations by considering the hypermatrix  $\mathbf{B}$  as independent of time. Let us begin by rewriting (31) as follows:

$$P = \frac{GM}{a_0} \sum_n \sum_m \sum_p \sum_q \left(\frac{a}{a_0}\right)^n F_{nmp} G_{npq} S_{nmpq} = \sum_{n,m,p,q} P_{nmpq}. \quad (23.42)$$

Clearly, each component  $P_{nmpq}$  can be treated separately and we get

$$\dot{\mathbf{k}} = \sum_{n,m,p,q} \mathbf{B} \nabla_k P_{nmpq} = \sum_{n,m,p,q} \dot{\mathbf{k}}_{nmpq}. \quad (23.43)$$

Each component  $P_{nmpq}$  thus contributes  $\dot{\mathbf{k}}_{nmpq}$  toward the total velocity  $\dot{\mathbf{k}}$ . Similarly, we can write

$$\delta\dot{\mathbf{k}}_{nmpq}(\tau) = \int_{\tau_0}^{\tau} \mathbf{B} \nabla_k P_{nmpq} d\tau. \quad (23.44)$$

Leaving the subscripts out for simplicity, the expressions for the individual contributions can now be derived.

The perturbation in  $a_0$  is given by the following formula:

$$\delta a_0(\tau) = \int_{\tau_0}^{\tau} \dot{a}_0 d\tau = \int_{\tau_0}^{\tau} (G M a_0)^{-1/2} 2 a_0 \frac{\partial P}{\partial \mu} d\tau \doteq 2 \sqrt{\frac{a_0}{GM}} \int_{\tau_0}^{\tau} \frac{\partial P}{\partial \mu} d\tau. \quad (23.45)$$

Writing  $\partial P / \partial \mu$  as  $(\partial P / \partial \psi)(\partial \psi / \partial \mu)$  and realizing that  $\partial \psi / \partial \mu = n - 2p + q$ , we get

$$\int_{\tau_0}^{\tau} \frac{\partial P}{\partial \mu} d\tau = (n - 2p + q) \int_{\tau_0}^{\tau} \frac{\partial P}{\partial \psi} d\tau. \quad (23.46)$$

Interchanging the variables, the last integral becomes

$$\int_{\tau_0}^{\tau} \frac{\partial P}{\partial \psi} d\tau = \int_{\tau_0}^{\tau} \dot{\psi}^{-1} dP, \quad (23.47)$$

and making use of the fact that

$$\dot{\psi} = (n - 2p) \dot{\omega} + (n - 2p + q) \dot{\mu} + m(\dot{\phi} - \dot{\theta}) \quad (23.48)$$

depends, in the first approximation, only on  $a_0, e, i$ , and  $J_2$ , we obtain

$$\delta a_0(\tau) \doteq \sqrt{\frac{a_0}{GM}} \frac{P}{\dot{\psi}} 2(n - 2p + q). \quad (23.49)$$

Substitution for  $P$  finally gives

$$\delta a_0(\tau) \doteq 2\sqrt{\frac{a_0}{GM}} \sum_{n,m,p,q} \left( \frac{a}{a_0} \right)^n \frac{F_{nmp}(i)G_{npq}(e)(n-2p+q)}{\dot{\psi}} S_{nmpq}(\psi). \quad (23.50)$$

A similar derivation for the remaining five orbital parameters yields [KAULA, 1966]

$$\begin{aligned} \delta e(\tau) &\doteq \sqrt{\frac{a_0}{GM}} \frac{\nu}{a_0 e} \sum_{n,m,p,q} \left( \frac{a}{a_0} \right)^n \\ &\quad \times \frac{F_{nmp}(i)G_{mpq}(e)[\nu(n-2p+q)-(n-2p)]}{\dot{\psi}} S_{nmpq}(\psi), \\ \delta i(\tau) &\doteq \sqrt{\frac{a_0}{GM}} \frac{1}{a_0 \nu \sin i} \sum_{n,m,p,q} \left( \frac{a}{a_0} \right)^n \\ &\quad \times \frac{F_{nmp}(i)G_{npq}(e)[(n-2p)\cos i - m]}{\dot{\psi}} S_{nmpq}(\psi), \\ \delta \mu(\tau) &\doteq \sqrt{\frac{GM}{a_0^3}} \left[ \tau - \sum_{n,m,p,q} \left( \frac{a}{a_0} \right)^n \right. \\ &\quad \times \left. \frac{F_{nmp}(i)[2(n+1)G_{npq}(e) - (\nu^2/e)G'_{npq}(e)]}{\dot{\psi}} \bar{S}_{nmpq}(\psi) \right], \quad (23.51) \\ \delta \varpi(\tau) &\doteq \sqrt{\frac{GM}{a_0^3}} \sum_{n,m,p,q} \left( \frac{a}{a_0} \right)^n \\ &\quad \times \frac{(\nu/e)F_{nmp}(i)G'_{npq}(e) - (1/\nu)\cot i F'_{nmp}(i)G_{npq}(e)}{\dot{\psi}} \bar{S}_{nmpq}(\psi), \\ \delta \omega(\tau) &\doteq \sqrt{\frac{GM}{a_0^3}} \frac{1}{\nu \sin i} \sum_{n,m,p,q} \left( \frac{a}{a_0} \right)^n \frac{F'_{nmp}(i)G_{npq}(e)}{\dot{\psi}} \bar{S}_{nmpq}(\psi), \end{aligned}$$

where  $F' = dF/di$ ,  $G' = dG/de$ , and (cf. (32))

$$\bar{S}_{nmpq}(\psi) = \begin{cases} -J_{nm}\sin \psi + K_{nm}\cos \psi, & n-m \text{ even}, \\ +J_{nm}\cos \psi + K_{nm}\sin \psi, & n-m \text{ odd}. \end{cases} \quad (23.52)$$

Since the system of perturbation equations (51) is not very clear, let us summarize them as follows:

$$\begin{aligned} \delta \tilde{k}_j(\tau) &\doteq \sum_{n=2}^{\infty} \sum_{m=0}^n \left[ J_{nm} \sum_{p=0}^n \sum_{q=-\infty}^{\infty} A_{j,nmpq}(a_0, e, i) \sigma(\psi) \right. \\ &\quad \left. + K_{nm} \sum_{p=0}^n \sum_{q=-\infty}^{\infty} A_{j,nmpq}(a_0, e, i) \bar{\sigma}(\psi) \right], j=1, \dots, 6, \end{aligned} \quad (23.53)$$

where  $A_{j,nmpq}$  are different functions of  $a_0, e, i, GM$ , and  $a$  for different orbital parameters, and  $\sigma, \bar{\sigma}$  are cosine or sine functions of  $\psi$  taken with the appropriate signs divided by  $\dot{\psi}$ . These equations are usually called equations of *linear perturbations*. They should be corrected for the effect of time dependence of  $P_T$  and  $B$ , as well as for the effect of the error in  $J_2$  evaluated from secular perturbations. The derivation of the corrections is very tedious, and we shall not treat it here. Let it suffice to state that these corrections alter only the functions  $A$  but do not affect the general form of (53).

### 23.4. Evaluation of gravity field parameters

It can be shown that the functions  $A_{j,nmpq}$  are approximately proportional to  $e^{|q|}$  [GAPOSHKIN AND LAMBECK, 1970]. For a typical geodetic satellite ( $e \sim 10^{-2}$ ), the magnitude of  $A$  decreases with the increasing absolute value of  $q$ . Hence the summation over  $q$  in (53) does not have to be carried too far, in the first approximation,  $q=0$  suffices. Also,  $A$  depends on time only through  $a_0, e, i$ . Since these vary with time very slowly, then  $A$  is approximately constant even for very long orbital arcs. The time dependence of  $\delta\tilde{k}$  is expressed basically through the trigonometrical terms  $\sigma, \bar{\sigma}$ , whose arguments  $\mu, \varpi, \omega$ , and  $\theta$  (cf. (33)) change with time in a more or less linear fashion. Clearly, we thus have the  $\dot{\psi}$  governing the periodicity of the perturbations. In  $\dot{\psi}$ ,

$$\dot{\mu} > \dot{\omega} - \dot{\theta} > \dot{\varpi}, \quad (23.54)$$

because  $\dot{\mu}$  is the orbital frequency of the satellite, typically several times per day,  $\dot{\omega} - \dot{\theta}$  is of the order of the frequency of the earth's spin, i.e., once a day since  $\dot{\theta} \gg \dot{\omega}$ ; and  $\dot{\varpi}$  is the frequency of the satellite perigee motion, generally much smaller than the other two frequencies. This ordering gives rise to the following terminology:

- (a) *short periodic variations* in  $\delta\tilde{k}$  are said to correspond to the frequencies of the order of once a day and higher, i.e.,  $> \dot{\theta}$ ;
- (b) *long periodic variations* represent frequencies between  $\dot{\theta}$  and  $\dot{\varpi}$ ; and
- (c) *secular periodic variations*, as we have already seen in §23.3, depict the changes in  $A$ .

It is easily seen from (48) that only for the combinations,

$$(n, m, p, q) = (2k, 0, k, 0), \quad k = 1, 2, \dots, \quad (23.55)$$

are all the coefficients by  $\varpi, \dot{\mu}$  and  $\dot{\omega} - \dot{\theta}$  zero. Thus only for these combinations are the corresponding partial perturbations varying purely secularly. A closer look leads to the discovery that the combinations (55) belong to the even-order zonal harmonics ( $m=0, n$  even) shown in §20.3 to be related to the ellipticity of the earth. Since the prevalent part (that for  $q=0$ ) of the perturbations, that due to the even-order zonal harmonics, possesses  $\dot{\psi}=0$ , it cannot be analysed using (53) because the coefficients by  $J_{nm}$  and  $K_{nm}$  become undefined. Linear equations similar to (41) must be used for this purpose. The odd-order zonal harmonics ( $m=0, n$  odd) and the tesseral harmonics ( $m \neq 0$ ) always cause short periodic perturbations.

We can clearly see that for some combinations of  $n, m, p, q$ , the factor  $\psi$  becomes small and the coefficients by  $J_{nm}, K_{nm}$  become magnified. Consequently, the particular frequencies for which  $\psi$  become small are called *resonant frequencies*; they are particularly useful in evaluating the corresponding potential coefficients. The strongest response of the satellite to the gravitational field occurs for the resonant frequencies, i.e., every satellite is especially sensitive to its own resonant frequencies.

Having had a closer look at the perturbation equations, we can now outline the approach to the solution for the potential coefficients. First, the bulk of the value of  $J_2$  is determined from secular perturbations. Next, the first approximation to the even-order, zonal coefficients and the correction to  $J_2$  are solved for once more using the secular perturbations. The last step consists of the temporal frequency decomposition (harmonic analysis according to  $\psi$ ) of the remaining perturbations. Evidently, always a whole set of potential coefficients is connected to a certain temporal frequency, and conversely, a whole series of temporal periods is related to each coefficient. Therefore, the amplitudes of individual periodic constituents, obtained from the harmonic analysis of the perturbations, correspond to different linear combinations of potential coefficients. Normally, perturbations of orbits of several different satellites are analysed simultaneously. In such a case, the linear combinations can be disentangled and the potential coefficients derived from a system of independent linear algebraic equations. One such solution is described in an illustrative way by GAPOSHKIN AND LAMBECK [1970]. It should be understood that the non-gravitational effects, if they are not corrected for prior to the harmonic analysis, may be solved for together with the potential coefficients.

How many potential coefficients can we solve for from the perturbations? Not very many; with the exception of coefficients corresponding to the resonant frequencies, the limit for  $n$  is about twenty. The reason for this rather low limit can be found in the *geometrical attenuation factor*  $(a/r)^n$  (cf. (30)) that regulates the upward propagation of the gravitational field. It very effectively filters out higher spatial frequency features with increasing altitude, as was pointed out in §23.1. To illustrate the power of this attenuation, let us take even a very low orbiting satellite at an altitude of 300 km. For this altitude, the attenuation factor is approximately equal to

$$\left(\frac{a}{r}\right)^n \doteq (1 - 0.05)^n, \quad (23.56)$$

and for  $n = 20$ , i.e., a feature wavelength of the order of 2000 km, the relative amplitude of the feature at the satellite altitude is only about 40% of that on the earth's surface (see FIG. 4).

Once the potential coefficients are known, it becomes a simple matter to determine other parameters of the gravity field. The perturbing gravitational potential (30), defined by these coefficients, is first converted to disturbing potential  $T$  using (20.89). Values of  $GM$ ,  $a$ , and  $J_n^N$ ,  $n = 2, 4, \dots$ , are selected according to the reference ellipsoid used. Then, using (20.90) and (22.12), we get

$$T_n \doteq \frac{R}{n-1} \Delta g_n \quad (23.57)$$

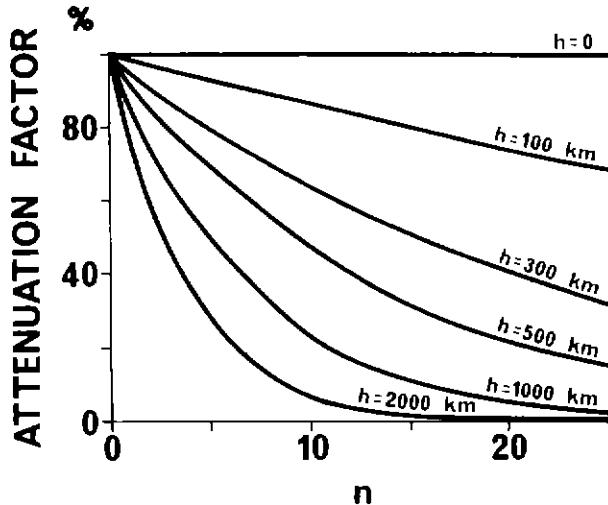


FIG. 23.4. Decrease of geometrical attenuation factor with altitude of orbit.

and

$$\Delta g \doteq \sum_{n=2}^{\infty} \frac{n-1}{R} T_n \quad (23.58)$$

valid on the reference ellipsoid. Substituting for  $T_n$ , we obtain

$$\Delta g \doteq -\frac{GM}{rR} \sum_{n=2}^l (n-1) \left(\frac{a}{r}\right)^n \sum_{m=0}^n [ (J_{nm} - J_{nm}^N) \cos m\lambda + K_{nm} \sin m\lambda ] P_{nm}(\sin \phi), \quad (23.59)$$

where  $l$  is the highest degree of the coefficients determined, and  $J_h^N$  for  $n = 3, 5, \dots$ , and  $J_{nm}^N$  for  $m \neq 0$  are all equal to zero (cf. §20.3).

On the earth's surface, the spherical approximation  $r \doteq a \doteq R$  can be used yielding

$$\Delta g \doteq \frac{GM}{R^2} \sum_{n=2}^l (n-1) \sum_{m=0}^n [ (J_{nm} - J_{nm}^N) \cos m\lambda + K_{nm} \sin m\lambda ] P_{nm}(\sin \phi). \quad (23.60)$$

This approximation introduces a maximum error of about 1 mGal [RAPP, 1972], which compares favourably with the currently obtainable standard deviation of the so-derived  $\Delta g$  that is of the order of 5 milligals. One such solution was shown in FIG. 6.9.

To derive an equation for geoidal height, an approach identical with that used in Chapter 22 is followed, i.e., the Bruns formula (21.4) is employed to convert the disturbing potential  $T$  to  $N$ . Using the same spherical approximation as above, we obtain

$$N \doteq -R \sum_{n=2}^l \sum_{m=0}^n [ (J_{nm} - J_{nm}^N) \cos m\lambda + K_{nm} \sin m\lambda ] P_{nm}(\sin \phi). \quad (23.61)$$

The maximum error in this approximation is around 1 m [RAPP, 1972] compared with the achievable standard deviation of  $N$ , which is of the order of 3 to 5 metres. Two examples of such satellite solutions, using two different reference ellipsoids, were seen in FIGS. 7.20 and 7.21.

The deflection components are obtained from  $N$  using (21.18). We get

$$\begin{aligned}\xi &= \sum_{n=2}^l \sum_{m=0}^n [(J_{nm} - J_{nm}^N) \cos m\lambda + K_{nm} \sin m\lambda] \frac{\partial P_{nm}(\sin \phi)}{\partial \phi}, \\ \eta &= \sum_{n=2}^l \sum_{m=0}^n [-(J_{nm} - J_{nm}^N) \sin m\lambda + K_{nm} \cos m\lambda] \frac{mP_{nm}(\sin \phi)}{\cos \phi}.\end{aligned}\quad (23.62)$$

For an example of numerical results, see FIGS. 6.23 and 6.24.

Generally, it should be remembered that the satellite solution as described in this chapter—often called *satellite dynamic solution for potential coefficients*—gives results that are more globally homogeneous than the results derived from terrestrial gravity data (cf. Chapter 22). On the other hand, the satellite solutions do not have the same detail, i.e., they are more generalized. Therefore a combination of the two techniques, exploiting the advantages of both, is preferable to any one of them deployed separately; we shall talk about this alternative in the next chapter. This argument, however, does not apply to the satellite techniques for the determination of  $GM$  and  $J_2$ . Satellite derived values of these two parameters are the most accurate values available today.

Satellite determined geocentric positions can also be used for the evaluation of gravity field parameters. This approach, which may be called *satellite geometrical solution for the geoid*, determines the geoidal height  $N$  as the difference between geodetic height  $h$  and orthometric height  $H^O$ —cf. §16.4. The geoidal height can be similarly determined, even on the seas, as the difference between the sea level height (above the reference ellipsoid) derived from satellite radar altimetry (see §19.4) and sea surface topography (cf. §7.2). For details, see, e.g., RAPP [1979].

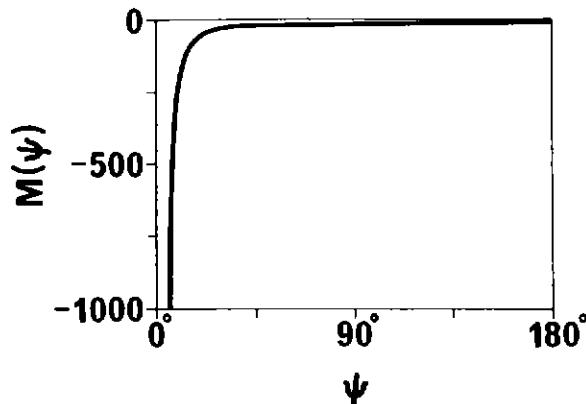
Once the geoidal heights are known, they can be converted to deflection components using Vening Meinesz's idea (see §22.1). Conversion of  $N$  to  $\Delta g$  is more involved. The expression relating  $\Delta g$  to  $N$  can be obtained by inverting Stokes's formula (22.17) by which process we get [MOLODENSKIY ET AL., 1960]

$$\Delta g(\phi_A, \lambda_A) = -\frac{\gamma N(\phi_A, \lambda_A)}{R} + \frac{\gamma}{4\pi R} \iint_{\mathbb{E}} M(\psi) [N(\phi, \lambda) - N(\phi_A, \lambda_A)] d\nu,\quad (23.63)$$

where the kernel is equal to

$$M(\psi) = -\frac{1}{4} \operatorname{cosec}^3 \frac{1}{2}\psi - 3 \cos \psi, \quad (23.64)$$

and  $\psi$  is the angular distance between  $(\phi_A, \lambda_A)$  and  $(\phi, \lambda)$ .

FIG. 23.5.  $M$  kernel.

Clearly, the conversion of  $N$  to  $\Delta g$  is a global problem, much the same as the conversion of  $\Delta g$  into  $N$ . Fortunately though, the kernel  $M$  tapers off very rapidly (cf. FIG. 5) so that it is not necessary to know the geoidal height differences all over the world. The gravity anomalies can be evaluated from geoidal heights confined to a reasonably small neighbourhood of the area of interest [COLEMAN AND MATHER, 1976]. On the other hand, the anomalies  $\Delta g$  vary much more rapidly than  $N$ , and the process thus is inherently unstable. This can be seen from the strong singularity of  $M(\psi)$  for  $\psi = 0$  compared with, e.g., Stokes's kernel.

A comparison of the satellite dynamics approach with the satellite techniques for positioning of a network of points (see §17.3) reveals that both mathematical models have one thing in common: they both use the observed orbit. Thus, under certain circumstances, it makes sense to combine the two tasks into one *combined positioning and potential coefficient determination*. Such combined solutions have been attempted and are described in, e.g., GAPOSHKIN [1973] or SEPPELIN [1974].

To conclude this chapter, let us at least mention one more concept the implementation of which is now being considered. It is *satellite gradiometry* whereby a three-dimensional accelerometer (see §16.1) is flown in the satellite. The satellite accelerations recorded by the accelerometer are then integrated with other kinds of orbital information to improve the knowledge of the orbit [CHOVITZ ET AL., 1973].

## CHAPTER 24

### DETERMINATION OF THE GRAVITY FIELD FROM DEFLECTIONS AND FROM HETEROGENEOUS DATA

In this chapter, the emphasis is put on the simple geometrical concept of geoidal height determination from the deflections of the vertical. As opposed to solutions described in the previous two chapters, this approach is of a local or regional character.

The first section is devoted to the geometrical solution for geoidal heights related to a geodetic rather than geocentric reference ellipsoid. The next section contains the transformations of gravity field parameters from one ellipsoid to another. It also touches on the problem of the determination of the best-fitting reference ellipsoid. In the third section, the possibilities of densifying the deflections of the vertical using different kinds of extraneous information are discussed. The last section contains the basic concepts of combined solutions for gravity field parameters from heterogeneous data.

#### 24.1. Geometrical solution for the geoid

It has already been shown in §21.1 that the geoidal deflection of the vertical is merely the maximum slope of the geoid with respect to the reference ellipsoid. We have also seen in §21.3 and §22.2 that the geoidal deflection can be obtained from the surface deflection by applying a correction for the effect caused by the curvature of the actual plumb line between the earth's surface and the geoid. In turn, the surface deflection can be derived from the geodetic and astronomical latitudes and longitudes of a common point (see (15.84) and (15.85)) and, as a result, in practice is often called the *astrodeflection*, or *astro-geodetic deflection*.

Thus, neglecting for the moment the curvature of the plumb line, the astronomical  $\Phi, \Lambda$  and geodetic coordinates  $\phi, \lambda$  of a point on the earth's surface give us all the necessary information about the geoidal slope at that point: the direction of the deflection coincides with the direction of the maximum slope (gradient), and the magnitude of the deflection equals that of the gradient (see FIG. 1). It should be evident that a slope specified in this manner is referred to the geodetic reference ellipsoid on which the geodetic coordinates are reckoned. Clearly, if the astronomical equal the geodetic coordinates, then the geoidal slope is equal to zero, and the geoid is, at that point, parallel to the geodetic reference ellipsoid.

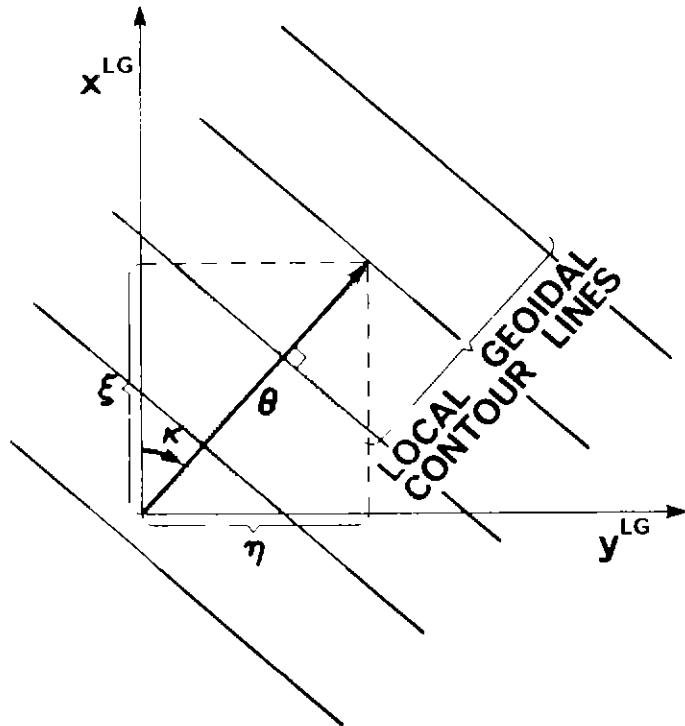


FIG. 24.1. Deflection components and geoidal slope.

As we have seen in §18.1, the reference ellipsoid needed for geodetic positioning is rarely geocentric. Therefore the geoid determined from astrodeflections is rarely referred to a geocentric ellipsoid as the gravimetric (cf. Chapter 22) and satellite (cf. Chapter 23) solutions are. Also, because the relation between the deflection components and the geoidal height is strictly local, knowledge of deflections from only the region of interest is required to obtain the geoid in that region.

Only geoidal height differences may be obtained from deflections. The equation linking the slope  $\epsilon$  of the geoid in any direction (cf. FIG. 16.21) with the geoidal height increment  $dN$  is simply a generalization of (21.18). Conforming to the sign convention for  $\xi$  and  $\eta$ , we can write

$$\epsilon = -\frac{dN}{dS}. \quad (24.1)$$

The relation between  $\epsilon$  and the deflection components was given by (16.80).

Let us now suppose that the deflection components along a curve on the geoid are known. Then the geoidal height difference between the two end points  $A, B$  of this curve can be evaluated as

$$N_B - N_A = \int_A^B dN = - \int_A^B \epsilon dS = - \int_A^B (\xi \cos \alpha + \eta \sin \alpha) dS. \quad (24.2)$$

In practice, the deflection components are known only at some geodetic control points, referred to as the *deflection points*, and the best we can do is to join the two end points by a traverse containing the available deflection points, as shown in

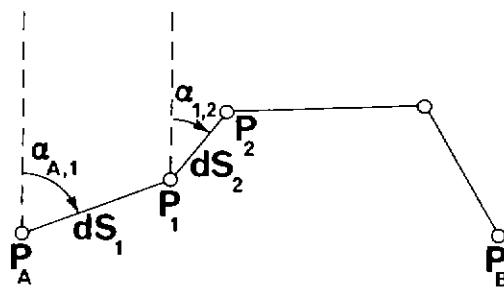


FIG. 24.2. Astro-geodetic levelling traverse.

FIG. 2. The geoidal height difference is then evaluated from the approximate formula

$$N_B - N_A \doteq - \sum_{i=1}^n \epsilon_i dS_i, \quad (24.3)$$

where we can take, for instance,

$$\epsilon_i = \frac{1}{2} [(\xi_{i-1} + \xi_i) \cos \alpha_{i-1,i} + (\eta_{i-1} + \eta_i) \sin \alpha_{i-1,i}]. \quad (24.4)$$

This technique was first proposed by HELMERT [1880] and became known as *astro-geodetic levelling*. Clearly, in a horizontal network, a number of different traverses may be selected that would follow different routes. These traverses can be joined to make a network, much the same way as levelling or gravity networks were made. The astro-geodetic levelling network can then be adjusted using techniques identical with those shown in §19.2.

To convert the adjusted geoidal height differences into geoidal heights, the geoidal height of at least one point, preferably a deflection point, must be known. It is natural to use the origin of the horizontal network for this purpose, since its geoidal height  $N_0$  is known by definition, being the quantity that fixes the separation of the reference ellipsoid from the geoid (cf. §18.1). However, any other point geoidal height may be used. An example of one such *astro-geodetic geoid* on the North American continent is shown in FIG. 3 [FISCHER ET AL., 1967]. This particular solution is referred to the NAD 27 (cf. §6.4).

It is easy to see that by utilizing the deflections of the vertical to obtain only the geoidal slope along a preconceived traverse, we do not use all the information inherent in the deflection components. It is clearly preferable to seek geoidal height variations in an areal manner through a two-dimensional approach. In such a two-dimensional approach, the geoidal height  $N$  is expressed as a function of position, say,  $x, y$ , where the coordinates  $x, y$  are taken in a local Cartesian system defined by (22.65). Geoidal height  $N$  can then be modelled as a two-dimensional linear form,

$$N(x, y) = \tilde{\Phi}^T(x, y)\mathbf{c} + c_0 = \tilde{N}(x, y) + c_0, \quad (24.5)$$

using any system  $\phi$  of  $n$  base functions. Since it is the slopes of the geoid in the

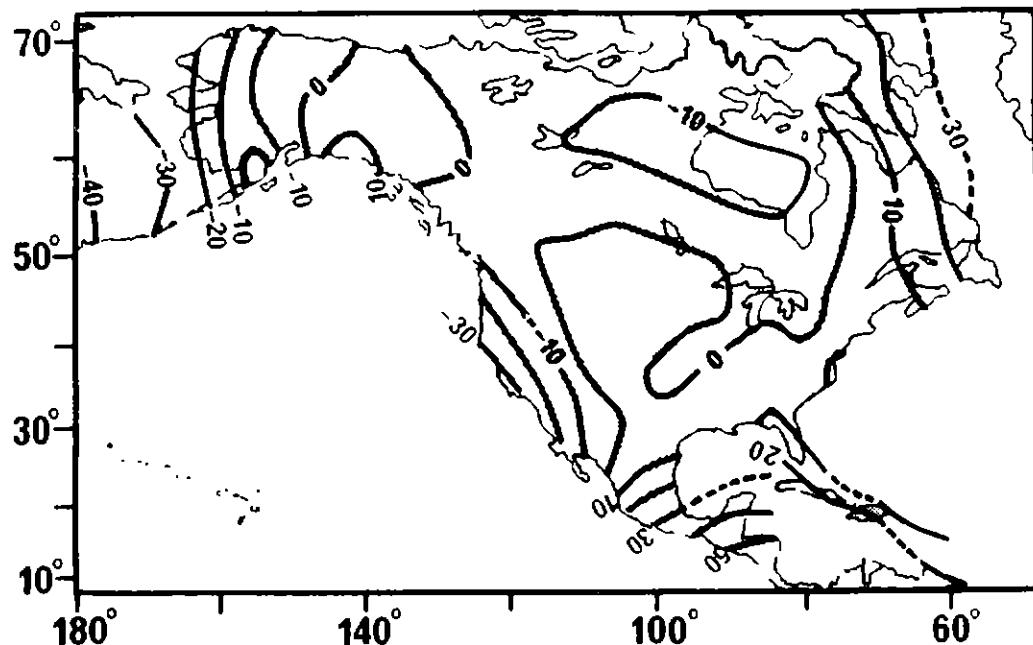


FIG. 24.3. U.S. Army Map Service geoid 1967 (referred to NAD 27). Contours in metres.

meridian,  $\xi$ , and the prime vertical,  $\eta$ , directions that are known, we write

$$\begin{aligned} -\xi(x, y) &= \frac{\partial N(x, y)}{\partial x} = \sum_{i=1}^n \frac{\partial \Phi_i(x, y)}{\partial x} c_i = \mathbf{B}_\xi^T(x, y) \mathbf{c}, \\ -\eta(x, y) &= \frac{\partial N(x, y)}{\partial y} = \sum_{i=1}^n \frac{\partial \Phi_i(x, y)}{\partial y} c_i = \mathbf{B}_\eta^T(x, y) \mathbf{c}. \end{aligned} \quad (24.6)$$

These are linear observation equations from which the coefficients  $\mathbf{c}$  can be evaluated using either a simple least-squares regression (§14.2) or collocation (§14.3). For example, the normal equations for the least-squares regression read

$$(\mathbf{B}_\xi \mathbf{C}_\xi^{-1} \mathbf{B}_\xi^T + \mathbf{B}_\eta \mathbf{C}_\eta^{-1} \mathbf{B}_\eta^T) \hat{\mathbf{c}} = -\mathbf{B}_\xi \mathbf{C}_\xi^{-1} \xi - \mathbf{B}_\eta \mathbf{C}_\eta^{-1} \eta. \quad (24.7)$$

As in the case of Helmert's astro-geodetic levelling, it is not possible to directly obtain the geoidal heights because of the nature of the observables, i.e., slopes. The computed surface  $\tilde{N}(x, y)$  is arbitrarily displaced—it has an arbitrary value  $\tilde{N}(0,0)$  at the origin of the coordinate system  $(0,0)$ . The absolute term  $c_0$  of the surface must be evaluated separately. The simplest way around this problem is to compute  $c_0$  from the following equations:

$$c_0 = N_0 - \tilde{N}(x_0, y_0), \quad (24.8)$$

where  $N_0 = N(x_0, y_0)$  is the defined value of the geoidal height at the initial point  $(x_0, y_0)$  of the network—not to be confused with  $N_0$  used in Chapter 22—and  $\tilde{N}(x_0, y_0)$  is the value computed from  $\tilde{\Phi}^T(x_0, y_0) \mathbf{c}$ .

The assembly of the appropriate covariance matrices  $\mathbf{C}_\xi$ ,  $\mathbf{C}_\eta$  is not trivial. Errors in both astronomical and geodetic coordinates (cf. §15.2 and §18.3) should be considered. In North America, the standard deviations of the astrodeflection components are of the order of  $1''$  for first-order and about  $1.5''$  for second-order deflection points. The values of the standard deviations increase with distance from the initial point of the network as the accuracy in geodetic positions  $\phi, \lambda$  decreases (see (7.1)). The main contribution to individual covariances comes from the covariance of geodetic coordinates that can be obtained from the covariance matrix of the adjusted geodetic coordinates.

As an example of such a two-dimensional solution, we may cite FIG. 7.17. It depicts the geoid related to the NAD 27 computed through least-squares regression using a mixed algebraic polynomial of order 8 ( $n = 63$ ). This solution may also serve as an example of how astro-geodetic determination of the geoid deteriorates with distance from the initial point of the network. According to (12.36), the covariance matrix  $\mathbf{C}_{\hat{c}}$  of the estimated geoidal coefficients  $c$  is given by

$$\boxed{\mathbf{C}_{\hat{c}} = (\mathbf{B}_\xi \mathbf{C}_\xi^{-1} \mathbf{B}_\xi^T + \mathbf{B}_\eta \mathbf{C}_\eta^{-1} \mathbf{B}_\eta^T)^{-1},} \quad (24.9)$$

and the standard deviation of the computed geoidal heights is (cf. (11.17))

$$\sigma_{\hat{N}}(x, y) = \sqrt{(\tilde{\Phi}^T(x, y) \mathbf{C}_{\hat{c}} \tilde{\Phi}(x, y))}. \quad (24.10)$$

The standard deviation of the aforementioned geoid is shown in FIG. 4 [VANÍČEK AND MERRY, 1973].

Evidently, using one approach or the other, the more astrodeflections we have in the region of interest, the better the solution. It can be shown that for the two-dimensional approach to give the same accuracy as astro-geodetic levelling, the deflection point spacing can be increased roughly twice [MERRY, 1975]. It should also be noted that the analytical, two-dimensional technique offers an additional advantage: once the coefficients  $c$  of the geoid are known, it becomes a simple matter to generate from (6) the deflection components at any point  $(x, y)$ . Nevertheless, this procedure has to be used with caution. The surface deflection field has more short wavelength features than the geoid because of the pronounced effect of topography on the surface deflections. The direct transformation  $(\xi, \eta \rightarrow N)$  then results in a smooth surface  $N$ , while the inverse procedure  $(N \rightarrow \xi, \eta)$  cannot regenerate the detail in the deflection field. This feature may be regarded as a smoothing of the *topographical noise in the deflections* and, as such, is not detrimental as long as the deflections are used for correcting horizontal angles in geodetic networks (see §18.3). It may not be so harmless, however, for other applications.

In practice, the correction to the astrodeflections for the curvature of the actual plumb line is seldom applied because its numerical evaluation is very laborious. The neglect of this correction results in the fact that the deflections being used are surface deflections  $\theta'$  (cf. §21.1) rather than geoidal deflections  $\theta$ . Since the surface deflections are numerically closer to the Molodenskij deflections than to the geoidal deflections, the surface computed from surface deflections is closer to the quasigeoid

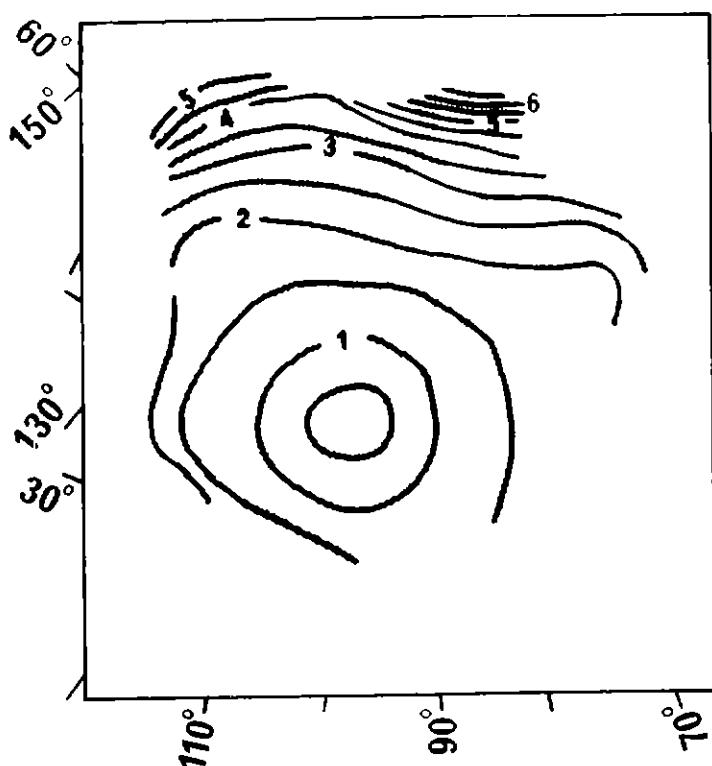


FIG. 24.4. Standard deviation of astro-geodetic geoid. Contours in metres.

than the geoid. It thus may deviate from the geoid significantly in mountainous regions, as we have seen in §22.2. In spite of this, the astro-geodetic solution at present offers the most accurate approximation to the actual geoid.

It should be stated that it is not usual in geodesy to transform the deflections into gravity anomalies. The reason is that gravity anomalies can be determined directly from gravity observations that are obtained more readily and much more cheaply than the deflections. If the need for doing this transformation should arise, the anomalies could be derived from astro-geodetically determined  $N$ , following the procedure, with all its pitfalls, described in §23.4.

## 24.2. Transformation of gravity field parameters

As we have seen in the previous section, the geoidal heights obtained from the astrodreflections are referred to the same geodetic reference ellipsoid as are the geodetic coordinates used in the derivation of the astrodreflections. On the other hand, the solutions described in Chapters 22 and 23 refer to a geocentric ellipsoid. To be able to compare the two kinds of solutions, we must be able to transform  $N$ , and similarly  $\xi$ ,  $\eta$ , from one ellipsoid to another. The transformation of  $\Delta g$  is normally not sought since it makes no sense to define a gravity anomaly referred to a geodetic ellipsoid.

To derive the transformation equations, let us begin by rewriting eqn. (15.97). Realizing that a change in latitude  $\delta\phi = \phi_2 - \phi_1$  is equivalent to a change in the meridian deflection component of  $-d\xi = \xi_1 - \xi_2$ ,  $\delta\lambda$  to  $-\delta\eta/\cos\phi$ , and  $\delta h$  to  $\delta N$ , we can write:

$$\begin{bmatrix} -\delta\xi \\ -\delta\eta/\cos\phi \\ \delta N \end{bmatrix} = -\mathbf{J}^{-1} \begin{bmatrix} \delta x_E \\ \delta y_E \\ \delta z_E \end{bmatrix} - \mathbf{J}^{-1}\mathbf{B} \begin{bmatrix} \delta a \\ \delta f \end{bmatrix} - \mathbf{J}^{-1}\mathbf{T} \begin{bmatrix} \delta\epsilon_x \\ \delta\epsilon_y \\ \delta\epsilon_z \end{bmatrix}. \quad (24.11)$$

It is then easy to multiply the first component by  $-1$  and the second by  $-\cos\phi$  to obtain immediately:

$$\boxed{\begin{bmatrix} \delta\xi \\ \delta\eta \\ \delta N \end{bmatrix} = \mathbf{D} \begin{bmatrix} \delta a \\ \delta f \end{bmatrix} + \mathbf{E} \begin{bmatrix} \delta x_E \\ \delta y_E \\ \delta z_E \end{bmatrix} + \mathbf{F} \begin{bmatrix} \delta\epsilon_x \\ \delta\epsilon_y \\ \delta\epsilon_z \end{bmatrix}}, \quad (24.12)$$

where, to the same degree of accuracy as  $\mathbf{B}$ ,  $\mathbf{J}^{-1}$ , and  $\mathbf{T}$  (cf. §15.4)

$$\mathbf{D} \doteq \begin{bmatrix} 0 & -2\sin\phi\cos\phi \\ 0 & 0 \\ -1 & a\sin^2\phi \end{bmatrix}, \quad (24.13)$$

$$\mathbf{E} \doteq \begin{bmatrix} -\sin\phi\cos\lambda/a & -\sin\phi\sin\lambda/a & \cos\phi/a \\ -\sin\lambda/a & \cos\lambda/a & 0 \\ -\cos\phi\cos\lambda & -\cos\phi\sin\lambda & -\sin\phi \end{bmatrix}, \quad (24.14)$$

and

$$\mathbf{F} \doteq \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ \sin\phi\cos\lambda & \sin\phi\sin\lambda & -\cos\phi \\ 0 & 0 & 0 \end{bmatrix}. \quad (24.15)$$

Analysing eqn. (12) we notice several things. First, the impact of the change  $\delta a$  in size of the reference ellipsoid is confined to geoidal heights only; the deflection components are not affected at all. Second, the effect of the misalignment difference  $\delta\epsilon_z$  along the  $z$ -axes is felt only by the  $\eta$  component. Last, but not least, the misalignment differences have no (first order) effect on geoidal heights. This fact has been known for some time [MERRY AND VANÍČEK, 1974b], and it is the property of geoidal heights being insensitive to misalignments that makes the *geoid matching* mentioned in §18.1 the most accurate technique for translation component determination. Let us explain the technique here.

Suppose that we know geoidal heights  $N_1(\phi_i, \lambda_i)$  referred to a geocentric ellipsoid ( $a_1, f_1$ ) for a number of points  $(\phi_i, \lambda_i)$ . These geoidal heights typically would have been obtained from either a gravimetric, a satellite, or a combined solution. Suppose further that at the same time we know geoidal heights  $N_2(\phi_i, \lambda_i)$ , for the same points as above, referred to a geodetic ellipsoid ( $a_2, f_2$ ), with translation components  $(x_{E2}, y_{E2}, z_{E2})$  and misalignment angles  $(\epsilon_{x2}, \epsilon_{y2}, \epsilon_{z2})$ . These geoidal

heights would have been derived from astro-geodetic deflections. Then, from eqn. (12), we get:

$$\begin{aligned}\delta N(\phi_i, \lambda_i) &= N_2(\phi_i, \lambda_i) - N_1(\phi_i, \lambda_i) \\ &= -\delta a + a_1 \sin^2 \phi_i \delta f - \cos \phi_i \cos \lambda_i \delta x_E - \cos \phi_i \sin \lambda_i \delta y_E \\ &\quad - \sin \phi_i \delta z_E \\ &= a_1 - a_2 - a_1 \sin^2 \phi_i (f_1 - f_2) - \cos \phi_i \cos \lambda_i x_{E2} \\ &\quad - \cos \phi_i \sin \lambda_i y_{E2} - \sin \phi_i z_{E2},\end{aligned}\tag{24.16}$$

where misalignment angles are conspicuously missing. Equation (16) is a linear ‘observation equation’ relating the unknown translation components to the known (‘observed’) geoidal height differences. A straightforward application of the least-squares technique then yields the best estimates of translation components.

It may happen in practice that the mutual position of the two datums, for which we seek the transformation of the deflections and geoidal heights, is known only at a common initial point  $(\phi_0, \lambda_0, h_0)$ . Let us assume, for simplicity, that each datum is positioned and oriented (with respect to the CT system) by the set of topocentric parameters  $\phi_0, \lambda_0, \alpha_0, \xi_0, \eta_0, h_0$ , of which the first three are common and only the last three differ by  $\delta \xi_0, \delta \eta_0, \delta N_0$ . Can the transformation equations still be formulated? It turns out that they can, but only if either the misalignment differences or the translation differences may be considered negligible.

To formulate the transformation equations, we proceed as follows. First, write eqn. (12) for the initial point. We get

$$\begin{bmatrix} \delta \xi_0 \\ \delta \eta_0 \\ \delta N_0 \end{bmatrix} = \mathbf{D}_0 \begin{bmatrix} \delta a \\ \delta f \end{bmatrix} + \mathbf{E}_0 \begin{bmatrix} \delta x_E \\ \delta y_E \\ \delta z_E \end{bmatrix} + \mathbf{F}_0 \begin{bmatrix} \delta \epsilon_x \\ \delta \epsilon_y \\ \delta \epsilon_z \end{bmatrix},\tag{24.17}$$

where the subscript 0 means that the matrices have to be evaluated specifically for  $(\phi_0, \lambda_0)$ . We can now premultiply eqn. (12) by  $\mathbf{E}^{-1}$  and eqn. (17) by  $\mathbf{E}_0^{-1}$ , subtract eqn. (17) from eqn. (12), and premultiply the result by  $\mathbf{E}$ . We obtain

$$\begin{bmatrix} \delta \xi \\ \delta \eta \\ \delta N \end{bmatrix} = \mathbf{E} \mathbf{E}_0^{-1} \begin{bmatrix} \delta \xi_0 \\ \delta \eta_0 \\ \delta N_0 \end{bmatrix} + (\mathbf{D} - \mathbf{E} \mathbf{E}_0^{-1} \mathbf{D}_0) \begin{bmatrix} \delta a \\ \delta f \end{bmatrix} + (\mathbf{F} - \mathbf{E} \mathbf{E}_0^{-1} \mathbf{F}_0) \begin{bmatrix} \delta \epsilon_x \\ \delta \epsilon_y \\ \delta \epsilon_z \end{bmatrix},\tag{24.18}$$

where everything is known except the misalignment differences, and the equation can be used only if these may be considered insignificant. It is left to the reader to verify that the following equation is also valid:

$$\begin{bmatrix} \delta \xi \\ \delta \eta \\ \delta N \end{bmatrix} = \mathbf{F} \mathbf{F}_0^{-1} \begin{bmatrix} \delta \xi_0 \\ \delta \eta_0 \\ \delta N_0 \end{bmatrix} + (\mathbf{D} - \mathbf{F} \mathbf{F}_0^{-1} \mathbf{D}_0) \begin{bmatrix} \delta a \\ \delta f \end{bmatrix} + (\mathbf{E} - \mathbf{F} \mathbf{F}_0^{-1} \mathbf{E}_0) \begin{bmatrix} \delta x_E \\ \delta y_E \\ \delta z_E \end{bmatrix}.\tag{24.19}$$

This equation can be used only if the translation differences are considered insignificant.

To complete this section, we shall show how gravity field parameters are used in determining the parameters (cf. §15.4) of the best-fitting horizontal datum. Even though reasons for wanting to use the best-fitting ellipsoid have mostly vanished, its determination constitutes an illustrative problem. The best-fitting ellipsoid can be sought in two natural ways—to render as minimum either the geoidal heights (squared) or the deflections of the vertical (squared) in the region of interest. Let us focus our attention here on the first approach which requires that the condition

$$\min_{a, f, x_E, y_E, z_E} \sum_{i=1}^n N^2(\phi_i, \lambda_i) \quad (24.20)$$

be satisfied. Denoting the parameters of the existing ellipsoid by subscript 1, we obtain from (16)

$$\begin{aligned} N(\phi_i, \lambda_i) = & N_1(\phi_i, \lambda_i) + a_1 - a - a_1 \sin^2 \phi_i (f_1 - f) \\ & + \cos \phi_i \cos \lambda_i (x_{E1} - x_E) \\ & + \cos \phi_i \sin \lambda_i (y_{E1} - y_E) + \sin \phi_i (z_{E1} - z_E), \quad i = 1, \dots, n. \end{aligned} \quad (24.21)$$

These linear observation equations, together with condition (20), yield the least-squares estimates of the best-fitting ellipsoidal parameters  $a, f, x_E, y_E, z_E$ .

Clearly, if the deflections are used instead of geoidal heights, similar observation equations are obtained, while the minimum condition will be given by (7.20). A problem arises, however, with the major semi-axis  $a$  that cannot be determined. It should also be pointed out that unlike the case of the (global) geocentric best-fitting ellipsoid (cf. §7.3), the two minimum conditions here will give different sets of parameters; hence the two best-fitting geodetic ellipsoids will generally be different.

The best-fitting horizontal datum may also be sought in terms of the topocentric position parameters (cf. §15.4), i.e., in terms of  $a, f, \xi_0, \eta_0, N_0$ . In such a case, (18) or (19) would be used. A strong correlation of  $a$  with  $N_0$  can be expected in this approach. The determination of the best-fitting geocentric ellipsoid does not differ conceptually from that of the geodetic ellipsoid. However, since the flattening is known to a high degree of accuracy from satellite tracking and the translation components are equal to zero by definition, the major semi-axis  $a$  is the only parameter sought.

In closing this section, it should be reiterated that it is impractical to adjust the misalignment of the geodetic ellipsoid with, say, the CT coordinate system using the gravity field parameters because they are insensitive to small angular displacements. We have seen this property earlier; it can be readily understood when we realize that the ellipsoid with the appropriate flattening is very close to a sphere. The misalignment angles are best determined using techniques discussed in §17.4.

### 24.3. Densification and refinement of deflections of the vertical

In §22.3, it was observed that for many applications, it is necessary to densify the gravity coverage. The same is true for the deflections of the vertical. In many areas,

either the deflection points are too widely separated geographically or the deflections are not known accurately enough to be useful. In this section, we shall show the different concepts used in the *prediction of deflections* by means of auxiliary information.

Let us focus our attention first on the possibility of *densification of observed astrodeflections*. The most obvious auxiliary data that can be used for this task are gravity anomalies: they are the cheapest, ‘directly observable’ parameters of the gravity field and are thus often available in adequate quantities. For any point, as we have seen, the geoidal deflection components, referred to a geocentric ellipsoid, can be determined from eqns. (22.24), which require the knowledge of gravity anomalies from all over the world. These deflections, which we shall call *gravimetrical deflections*, may then be converted to deflections referred to the geodetic reference ellipsoid by means of (12), (18) or (19). These, in turn, may then be corrected for the curvature of the actual plumb line to give surface deflections compatible with astrodeflections.

The problems with this approach are many. First, the gravimetrical deflections are not very accurate; although their local variations may be quite realistic, the long wavelength features are usually biased because of the non-homogeneous global gravity coverage. Second, the mutual position of the two reference ellipsoids involved may not be very well known. Third, the evaluation of the gravimetrical deflections requires the application of the correction for the indirect effect (cf. §22.1). Finally, the calculation of the curvature correction is always tedious. For these reasons, MOLODENSKIJ ET AL. [1960] proposed a somewhat simpler alternative based on the following assessment: when gravimetrical deflections in a small region are computed, the effect of distant gravity anomalies varies quite slowly from point to point. Hence, if the Vening Meinesz integration (22.24) is carried out over a sufficiently large neighbourhood (a few hundred kilometres) of the region of interest, we obtain *incomplete gravimetrical deflections* that differ from the correct ones by an almost constant amount. Similarly, for a small region, the corrections to gravimetrical deflections due to the difference of the two ellipsoids, and due to the indirect effect are practically constant.

Between any two adjacent deflection points, the conglomerate of the three effects (biases) can therefore be treated as varying linearly, and the situation depicted in FIG. 5 occurs. The reader can easily verify that the linear relation between the projected incomplete gravimetrical deflection  $\epsilon'$ , and the projected surface deflection  $\epsilon'$ , both taken in the vertical plane containing the two deflection points  $P_A$ ,  $P_B$ , is

$$\epsilon' = \epsilon_i - (\epsilon_{iA} - \epsilon'_A) + \frac{S}{S_{AB}} (\epsilon_{iA} - \epsilon'_A - \epsilon_{iB} + \epsilon'_B). \quad (24.22)$$

In order to predict the surface deflection  $\epsilon'$  from the above formula, the astrodeflections  $\epsilon'_A$ ,  $\epsilon'_B$  have to be known.

Clearly, the described prediction along the straight line connecting the two deflection points suffers from the same reduction of dimensionality as does Helmert’s astro-geodetic levelling. It is again preferable to utilize the inherent two-dimensionality of both kinds of deflections by modelling their differences by a plane or a low

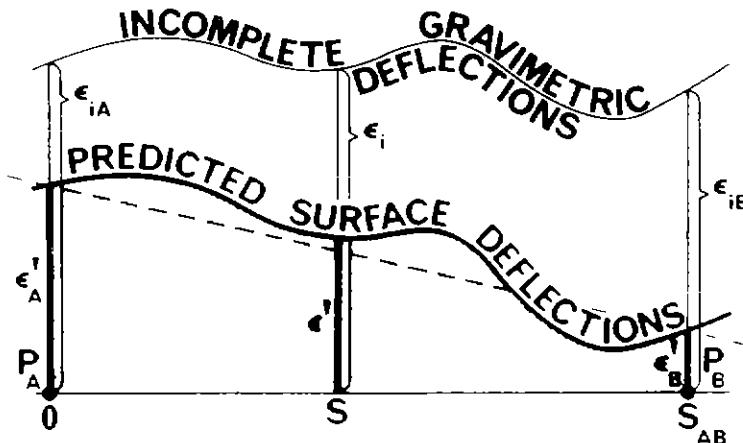


FIG. 24.5. Prediction of surface deflections.

order (smooth) surface. Such a *difference surface* is, of course, used for an area comprising several deflection points, rather than just for the line connecting the two deflection points. The situation is shown diagrammatically in FIG. 6, where  $\theta'$  denotes the astrodeflections (perhaps reduced to the geoid by curvature corrections), and  $\theta_i$  denotes the incomplete gravimetric deflections. Contour lines of the difference surface are dashed.

A search for the difference surface leads to regression equations similar to (6) with the only exception being that instead of  $\xi, \eta$ , the differences  $\xi' - \xi$ , and  $\eta' - \eta$ , are used. An example of results from this two-dimensional approach using simple regression was seen in FIG. 6.22 where some of the deflections, notably those in the sea, were predicted. It should be noted that in this particular example, the astrodeflections were not reduced to the geoid; the relatively low altitudes of the deflection points did not seem to warrant the effort. Even then, the accuracy of deflections predicted in this way is fairly good. If the astrodeflection coverage is adequately dense (i.e., with spacing of no more than some 50 km), then the standard deviations of the predicted deflections are only slightly larger than those of the original astrodeflections [MERRY AND VANÍČEK 1974c]. The technique of least-squares collocation can be used instead of or in addition to the regression [TSCHERNING AND FORSBERG, 1978] giving good results.

It should be mentioned that since surface deflections are correlated to a certain extent with shallow depth mass density anomalies (cf. §6.4), i.e., mainly with the terrain relief, we can also use the knowledge of topography and subsurface density to predict local variations of the deflections. Attempts in this direction can be found in, e.g., HELMERT [1900] and FISCHER [1974]. Also, inertial positioning devices used in a particular operational mode are capable of measuring the variations in the surface deflection components along the route joining two deflection points. The accuracy of deflections densified in this way is of the order of  $1''$  [GREGERSON, 1980].

A cheap source of geoidal deflection data is the satellite derived potential coefficients (see (23.62)) corrected for the gravitational effect of the atmosphere (see

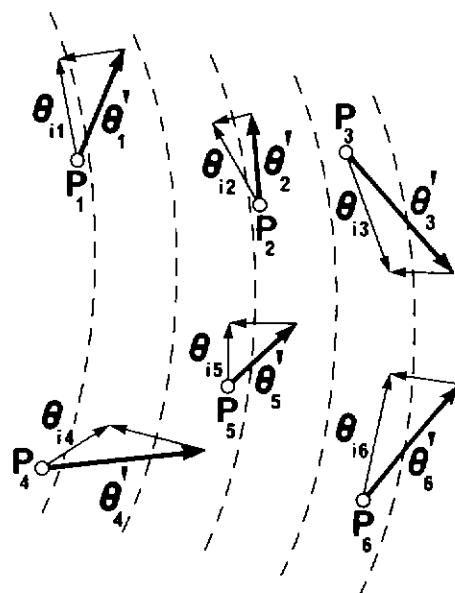


FIG. 24.6. Differences between astrodeflections and incomplete gravimetric deflections.

§9.4). Deflections determined in this manner are, as we have seen in Chapter 23, too smooth to be of much use in geodetic applications. Once again, terrestrial gravity anomalies may be used to provide the detailed structure of the field. There exist several techniques for such *refinement of the geoidal deflections*:

(a) First, the satellite derived geoidal deflection can be evaluated on a mesh of appropriately spaced points. These deflections represent the long wavelength contribution and play the same role as the astrodeflections did in the densification of surface deflections described above. Consequently, from this stage on, the technique is identical with the aforementioned densification of surface deflections.

(b) Alternatively, the low order, satellite derived potential coefficients, corrected for the atmospheric effect, can be merged with the higher order potential coefficients derived from gravity anomalies (cf. (22.8)). Because of the global orthogonality of spherical harmonics (cf. (20.37)), the two sets of potential coefficients are statistically independent, and the two sets can be treated together provided the two solutions are indeed referred to the same geocentric ellipsoid. The resulting set of potential coefficients is then converted into geoidal deflections by means of (23.62); the situation is shown in FIG. 7.

(c) An interesting alternative has been developed by LACHAPELLE [1978], who uses the satellite determined potential coefficients to compute the long wavelength part of the geoidal deflections (23.62). This part is then combined with the short wavelength contribution evaluated from gravity anomalies—using Vening Meinesz's formula and least-squares collocation—and astrodeflections. The accuracy of these refined deflections is between  $1.0''$  and  $1.5''$ . More about this kind of approach will be shown in §24.4.

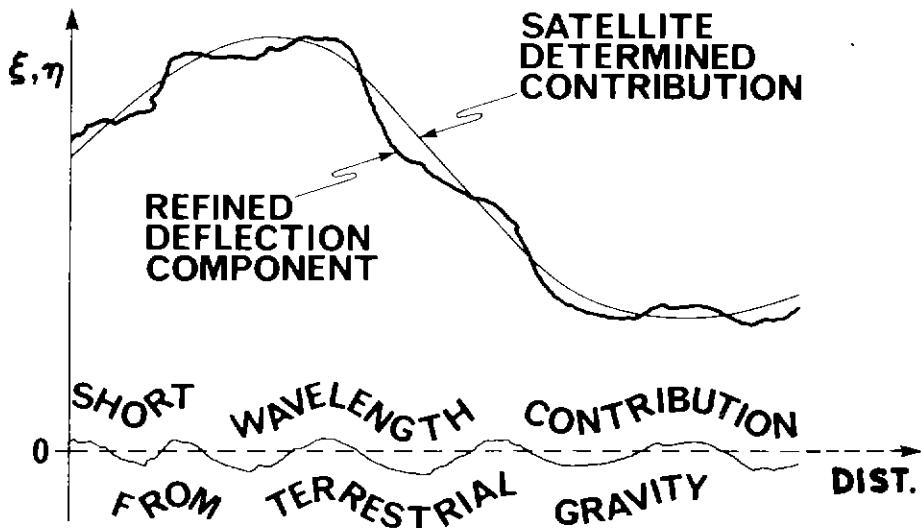


FIG. 24.7. Refinement of geoidal deflections.

#### 24.4. Solutions for the geoid from heterogeneous data

In the previous sections of this part, we have learned that the geoid (geoidal heights) can be either determined directly from observations (cf. §23.4) or derived from other gravity field parameters (from  $\Delta g$  in §22.1 and §22.2, from satellite derived potential coefficients in §23.4, and from  $\xi, \eta$  in §24.1). These determinations are not equivalent;

(a) We have seen that the gravimetric solution ( $\Delta g \rightarrow N$ ) provides a good local detail, but the long wavelength content is biased because of the irregular distribution of gravity observations on the earth's surface.

(b) The satellite solution ( $T \rightarrow N$ ) has unbiased and homogeneous long wavelength features but no detail.

(c) The astro-geodetic solution ( $\xi, \eta \rightarrow N$ ) is based on (deflection) data, of necessity, widely spaced, and so is the geoid obtained from directly measured geoidal heights. In addition, the accuracy of the astro-geodetic solution is not homogeneous; it deteriorates with distance from the initial point of the geodetic horizontal network.

These facts naturally suggest that combinations of the various kinds of data should be used, as we have already done in §24.3 for the case of the deflection components. By properly combining different kinds of data, and exploiting the better global homogeneity inherent in one data set on the one hand, and the higher local accuracy of the other data set on the other hand, one should be able to enhance the quality of the result. Clearly, there are numerous ways to properly combine heterogeneous data to advantage. It would be superfluous to even try to enumerate them all here; instead, we shall discuss only the concepts and describe some of the more popular possibilities.

When the geoid is needed purely for positioning (see §16.2), it is more straightforward to directly seek the geoidal heights referred to the local geodetic reference ellipsoid. In such a case, the astrodeflections and the directly determined geoidal

heights (above the geodetic reference ellipsoid) are the natural basic data to use for the task. These can be combined, for instance, with the densified deflections from gravity (cf. §24.3). In this combination, the long wavelength bias of a gravimetric solution is removed while the point values of astrodeflections are properly bridged. A geoid computed from these combined deflections along the lines shown in §24.1 is commonly known as an *astro-gravimetric geoid*. An example of such a geoid [MERRY AND VANÍČEK, 1974c] was shown in FIG. 6.15.

An astro-gravimetric geoid could be, of course, computed using different approaches. For instance, if the geoid is sought as a linear form (cf. (5)), then two sets of coefficients can be computed from the normal equations (7), obtained from the astrodeflections, and the normal equations (22.91), formulated for the gravity anomalies. The bias of the gravimetric or, more likely, a partial gravimetric solution, has to be removed using a difference surface similar to the one used for removing the bias in incomplete gravimetric deflections and described in §24.3. The effect of distant gravity anomalies on geoidal heights does not taper off as rapidly as the effect on the deflections. Thus the integration in Stokes's formula has to cover a wider neighbourhood of the area of interest than Vening Meinesz's formula. Thus the kernel in (22.92) should be based on Stokes's rather than Vening Meinesz's function.

It should be pointed out that if the geodetic reference ellipsoid is positioned within the earth through satellite determined coordinates of some control points (see §18.1), then it becomes more natural to use the directly determined geoidal heights at these points as the basic data. When these geoidal heights are selected for the basis, the systematic bias in the geoidal heights derived from other sources is mostly removed by making the latter fit the directly determined geoidal heights as well as possible. As a result, the remaining bias in the resulting geoidal heights becomes consistent with the bias in the position of the geodetic ellipsoid and no longer has an adverse effect on position computations. There is nothing, of course, to prevent us from employing additional kinds of data in computing a geoid related to the geodetic reference ellipsoid, if these are readily available.

When geoidal heights referred to a geocentric ellipsoid are sought, the most popular choice of data used for this task are the satellite derived potential coefficients (corrected again for the gravitational effect of the atmosphere) and terrestrial free air gravity anomalies. It is the same combination as that used for the evaluation of geoidal deflections of the vertical (cf. §24.3). In this approach, the corrected potential coefficients  $J'_{nm}(T)$ ,  $K'_{nm}(T)$  up to  $l$  degree are taken to define the *reference spheroid of degree  $l$*  (cf. §7.2). Its spheroidal heights  $N'$  are defined by (23.61). Then we can write

$$\begin{aligned} N(\phi, \lambda) &= -R \sum_{n=2}^l \sum_{m=0}^n [(J'_{nm}(T) - J_{nm}^N) \cos m\lambda + K'_{nm}(T) \sin m\lambda] P_{nm}(\sin \phi) \\ &\quad - R \sum_{n=l+1}^{\infty} \sum_{m=0}^n [(J_{nm}(T) - J_{nm}^N) \cos m\lambda + K_{nm}(T) \sin m\lambda] P_{nm}(\sin \phi) \\ &= N'(\phi, \lambda) + \delta N'(\phi, \lambda), \end{aligned} \quad (24.23)$$

where the potential coefficients  $J_{nm}(T)$ ,  $K_{nm}(T)$  from the degree  $l+1$  up are evaluated from the terrestrial gravity anomalies, using (22.8) and multiplying the results by  $-a/(GM)$ , see §20.3.

It is interesting to realize that as a consequence of the global orthogonality of spherical harmonics, the coefficients  $J_{nm}(\Delta g)$ ,  $K_{nm}(\Delta g)$  remain the same when any linear combination of terms  $(a_{ij} \cos j\lambda + b_{ij} \sin j\lambda)P_{ij}(\sin \phi)$ , for  $i \neq n$  and  $j \neq m$ , is added to  $\Delta g$ . This property can be used to get the smallest absolute values of gravity anomalies (modified in this way) to ease the numerical evaluation of the potential coefficients as much as possible. Clearly, the smallest absolute values result from subtracting the corrected, satellite determined, low order harmonic components of the anomaly field obtained from (23.60). We get

$$\begin{aligned}\delta\Delta g' &= \Delta g - \frac{GM}{R^2} \sum_{n=2}^l (n-1) \sum_{m=0}^n [ (J'_{nm}(T) - J^N_{nm}) \cos m\lambda \\ &\quad + K'_{nm}(T) \sin m\lambda ] P_{nm}(\sin \phi) \\ &= \Delta g - \Delta g'.\end{aligned}\tag{24.24}$$

For  $n > l$ , we have, of course,

$$J_{nm}(\Delta g) = J_{nm}(\delta\Delta g'), \quad K_{nm}(\Delta g) = K_{nm}(\delta\Delta g').\tag{24.25}$$

Often, it is desirable to express the geoidal height  $\delta N'$  (referred to the reference spheroid of degree  $l$ ) in a closed form. Transformation of the second series in (23) into a closed form can be done following the Stokes approach (see §22.1), and it results in

$$\delta N' \doteq \frac{R}{4\pi\gamma_0} \oint_S \Delta g(\phi, \lambda) \sum_{n=l+1}^{\infty} \frac{2n+1}{n-1} P_n(\cos \psi) d\nu.\tag{24.26}$$

First, we observe that because this formula is equivalent to the second series in (23), it must also yield the identical result  $\delta N'$  if  $\delta\Delta g'$  is substituted for  $\Delta g$ . Further, the series in (26) can also be written in a finite form (see (22.15)) as

$$S_l(\psi) = S(\psi) - \sum_{n=2}^l \frac{2n+1}{n-1} P_n(\cos \psi),\tag{24.27}$$

where  $S(\psi)$  is Stokes's function. We shall call  $S_l$  (of which, of course,  $S_2 = S$ ) the *spheroidal Stoke's function*. The formula for the geoidal height above the reference spheroid of degree  $l$  then reads

$$\delta N' \doteq \frac{R}{4\pi\gamma_0} \oint_S \delta\Delta g'(\phi, \lambda) S_l(\psi) d\nu.\tag{24.28}$$

Some spheroidal Stokes's functions are shown in FIG. 8, according to WONG AND GORE [1969]. A further treatment of these functions can be found in LACHAPELLE

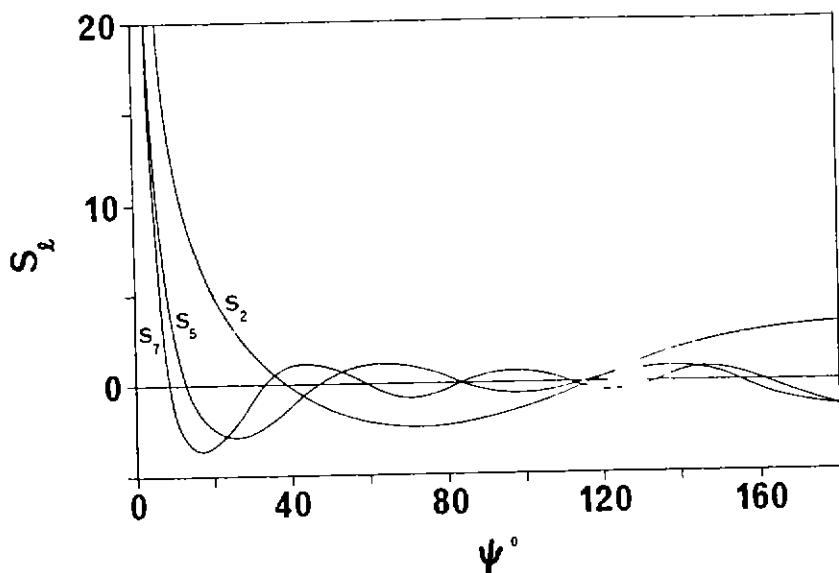


FIG. 24.8. Spheroidal Stokes's functions.

[1977]. Clearly, the spheroidal Stokes's functions converge toward zero more rapidly than the ordinary (ellipsoidal) Stokes's function: the higher the order of the reference spheroid, the faster the convergence. The significance of this faster

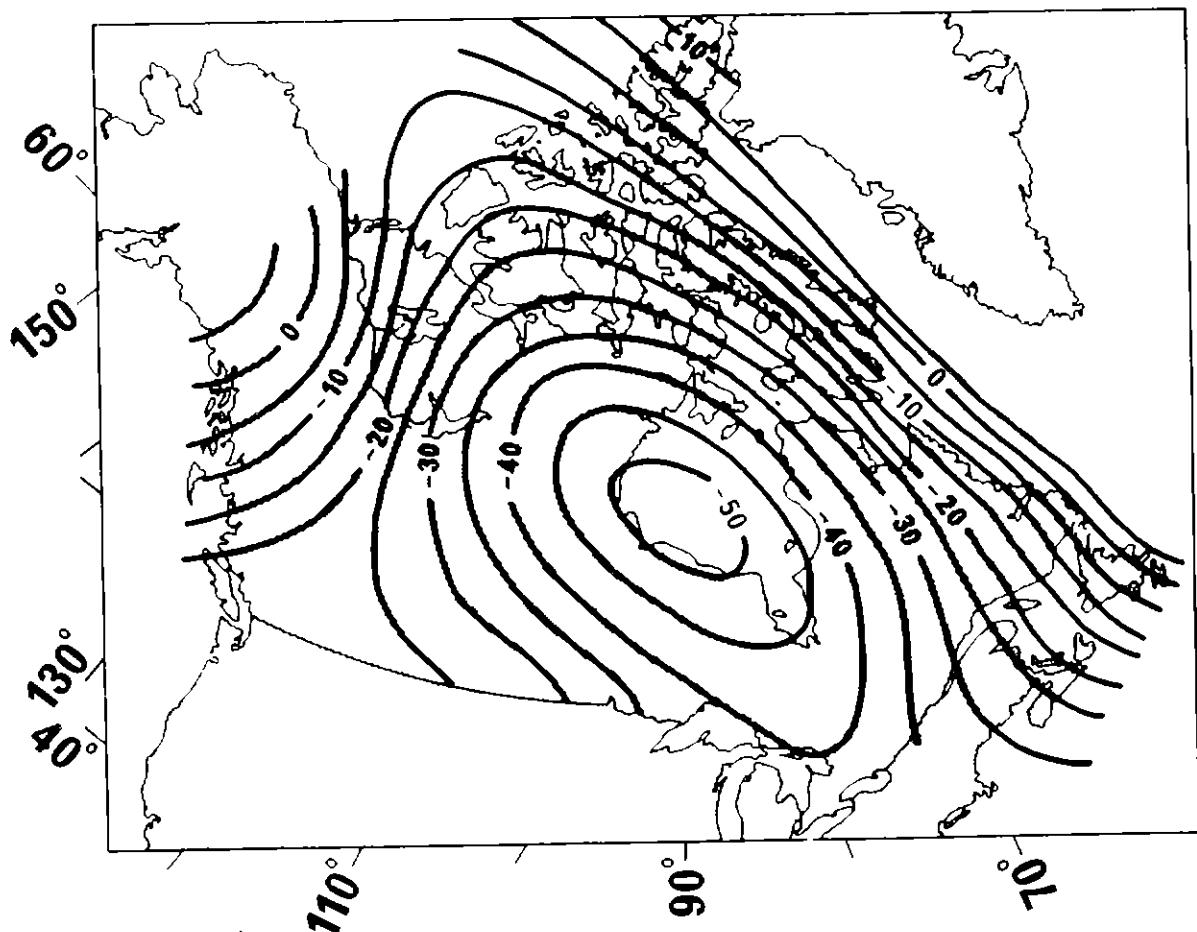


FIG. 24.9. GEM 10 in Canada. Contours in metres.

convergence is that the effect of distant anomalies diminishes with increasing  $l$ , and the integration (28) does not have to be carried out too far. This result only confirms the well-known behaviour of  $\delta N'$ . It should be noted that since the two components of the so-determined  $N$  are statistically independent, the variance of  $N$  is the sum of the variances in  $N'$  and  $\delta N'$ .

The indirect effect (cf. §22.1) of the free air anomalies used in these combined solutions should be accounted for. As stated in §24.3, the difference surface used in the first combination technique described above removes the bias arising from the indirect effect. In the latter combination technique, we have to worry only about the higher frequency ( $n > l$ ) components of the indirect effect, since the satellite solution is not affected. Fortunately, it appears that most of the indirect effect is concentrated in the low frequency band.

Naturally, other kinds of data can be used together with the satellite derived potential coefficients and terrestrial gravity. Mathematical models needed for such combinations are left to the reader to develop. Let us conclude this chapter by pointing out that we have already seen (FIG. 6.16) one solution for the geoid derived from satellite and terrestrial gravity data [VINCENT ET AL., 1972]. A more current solution, worked out at the Goddard Space Flight Center [LERCH ET AL., 1977] and known as GEM 10, is shown in FIG. 9, according to LACHAPELLE [1980]. It is referred to the mean earth ellipsoid of  $a = 6\,378\,135$  m and  $f = 1/298.257$ .

## PART V

### REFERENCES

- ABRAMOWITZ, M. AND I.A. STEGUN (EDS.) (1964). *Handbook of Mathematical Functions*. Dover reprint, 1965.
- BALMINO, G. (1972). Representation of the earth potential by buried masses. *Proc. 3rd International Symposium on the Use of Artificial Satellites for Geodesy*, Eds. S.W. Henriksen, A. Mancini and B.H. Chovitz. AGU, NOAA, Washington, D.C., U.S.A., April, 1971. American Geophysical Union Monograph 15, pp. 121–124.
- BJERHAMMAR, A. (1963). A new theory of gravimetric geodesy. Geodesy Division Report of the Royal Institute of Technology, Stockholm, Sweden.
- BODEMÜLLER, H. (1957). Beitrag zur Schwerekorrektion geometrischer Nivellements. Deutsche Geodätische Kommission, Reihe A, Höhere Geodäsie, Heft Nr. 26. Munich, Germany.
- BOMFORD, G. (1971). *Geodesy*. 3rd ed., Oxford University Press.
- BRUNS, H. (1878). *Die Figur der Erde*. Publication des Königlichen Preussischen Geodätischen Institutes, Berlin, Germany.
- CAPUTO, M. (1967). *The Gravity Field of the Earth*. Academic Press.
- CHOVITZ, B., J. LUCAS AND F. MORRISON (1973). Gravity gradients at satellite altitudes. NOAA Technical Report NOS 59, U.S. Department of Commerce, Rockville, U.S.A.
- COLEMAN, R. AND R.S. MATHER (1976). Computational procedures for the use of the inverse of Stokes' operator. Department of Geodesy, Unisurv No. G24, University of New South Wales, Sydney, Australia, pp. 123–139.
- DEPARTMENT OF ENERGY, MINES AND RESOURCES (1977). Personal communication. Gravity Division, Earth Physics Branch, Ottawa, Canada.
- DE SITTER, W. (1924). On the flattening and the constitution of the earth. *Bull. Astr. Inst. Neth.* 55.
- ESPOSITO, P.B. AND A.T.Y. NG (1976). Geocentric gravitational constant determined from spacecraft radiometric data. *Phys. of the Earth and Planetary Interiors* 12, pp. 283–289.
- FALLER, J.E. (1965). An absolute interferometric determination of the acceleration of gravity. *Bull. Géod.* 77, pp. 203–204.
- FISCHER, I. (1974). Deflections at sea. *J. Geophys. Res.* 79 (14), pp. 2123–2128.
- FISCHER, I., M. SLUTSKY, R. SHIRLEY AND P. WYATT (1967). Geoid charts of North and Central America. U.S. Army Map Service Technical Report 62, Washington, D.C., U.S.A.
- FORWARD, R.L. (1974). Review of artificial satellite gravity gradiometer techniques for geodesy. *Proc. International Symposium on the Use of Artificial Satellites for Geodesy and Geodynamics*, Ed. G. Veis. IAG, COSPAR, Athens, Greece, May, 1973. National Technical University, pp. 157–192.
- GAPOSHKIN, E.M. (1973). 1973 Smithsonian standard earth (III). Smithsonian Astrophysical Observatory Special Report 353, Cambridge, U.S.A.
- GAPOSHKIN, E.M. AND K. LAMBECK (1970). 1969 Smithsonian standard earth (II). Smithsonian Astrophysical Observatory Special Report 315, Cambridge, U.S.A.

- GOODKIND, J.M. (1978). High precision tide spectroscopy *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference, An International Symposium on the Applications of Geodesy and Geodynamics*, Ed. I.I. Mueller. IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280, The Ohio State University, Columbus, U.S.A., pp. 309–311.
- GREGERSON, L.F. (1979). Personal communication.
- GREGERSON, L.F. (1980). Personal communication.
- GROTN, E. AND H. MORITZ (1964). On the accuracy of geoid heights and deflections of the vertical. Department of Geodetic Science Report 38, The Ohio State University, Columbus, U.S.A.
- HAYFORD, J.F. AND W. BOWIE (1912). The effect of topography and isostatic compensation upon the intensity of gravity. U.S. Coast and Geodetic Survey Report 10, Washington, D.C., U.S.A.
- HEISKANEN, W.A. (1938). New isostatic tables for the reduction of the gravity values calculated on the basis of Airy's hypothesis. Publications of the Isostatic Institute of the IAG, No. 2, Helsinki, Finland.
- HEISKANEN, W.A. (1957). The Columbus geoid. *EOS, Trans. Am. Geophys. Union* 38 (6), pp. 841–848.
- HEISKANEN, W.A. AND H. MORITZ (1967). *Physical Geodesy*. Freeman.
- HEISKANEN, W.A. AND E. NISKANEN (1941). World maps for the indirect effect of the undulations of the geoid on gravity anomalies. Publications of the Isostatic Institute of the IAG, No. 7, Helsinki, Finland, reprinted from *Annales Academiae Scientiarum Fennicae Ser. A* 57 (4).
- HEISKANEN, W.A. AND F.A. VENING MEINESZ (1958). *The Earth and its Gravity Field*. McGraw-Hill.
- HELMERT, F.R. (1880). *Die mathematischen und physikalischen Theorien der höheren Geodäsie*. Vol. I., Minerva G.M.B.H. reprint, 1962.
- HELMERT, F.R. (1900). Zur Bestimmung kleiner Flächenstücke des Geoides aus Lotabweichungen mit Rücksicht auf Lotkrümmung. Sitzungsberichte Preuss. Akad. Wiss., Berlin, Germany.
- HIRVONEN, R.A. (1960). New theory of the gravimetric geodesy. Publications of the Isostatic Institute of the IAG, No. 32, Helsinki, Finland.
- HIRVONEN, R.A. (1962). On the statistical analysis of gravity anomalies. Publications of the Isostatic Institute of the IAG, No. 37, Helsinki, Finland.
- HOBSON, E.W. (1931). *The Theory of Spherical and Ellipsoidal Harmonics*. Cambridge University Press.
- HOCHSTADT, H. (1964). *Differential Equations*. Dover reprint, 1975.
- INTERNATIONAL ASSOCIATION OF GEODESY (1971). Geodetic Reference System, 1967. IAG Special Publication No. 3, Paris, France.
- INTERNATIONAL ASSOCIATION OF GEODESY (1974). The international gravity standardization net, 1971. Special Publication No. 4, Paris, France.
- INTERNATIONAL UNION OF GEODESY AND GEOPHYSICS (1976). *IUGG Chronicle*. No. 108, Paris, France.
- JEKELI, C. (1980). Reducing the error of geoid undulation computations by modifying Stokes's function. Department of Geodetic Science Report 301, The Ohio State University, Columbus, U.S.A.
- KAULA, W. (1959). Statistical and harmonic analysis of gravity. *J. Geophys. Res.* 64, pp. 2401–2421.
- KAULA, W. (1962). Celestial geodesy. In Vol. 9 of *Advances in Geophysics*, Eds. H.E. Landsberg and J. van Mieghem. Academic Press, pp. 192–293.
- KAULA, W. (1963). Determination of the earth's gravitational field. *Rev. Geophys. and Space Phys.* 1 (4), pp. 507–551.
- KAULA, W. (1966). *Theory of Satellite Geodesy: Applications of Satellites to Geodesy*. Blaisdell.
- KING-HELE, D.G. (ORGANIZER) (1967). A discussion on orbital analysis. *Philos. Trans. Roy. Soc. London Ser. A* 262.
- KOBOLD, F. AND E. HUNZIKER (1962). Communication sur la courbure de la verticale. *Bull. Géod.* 65, pp. 265–267.
- KOCH, K.R. AND F. MORRISON (1970). A simple layer model of the geopotential from a combination of satellite and gravity data. *J. Geophys. Res.* 75, pp. 1483–1492.
- KOUBA, J. (1979). Personal communication. Geodetic Survey of Canada, Department of Energy, Mines and Resources, Ottawa, Canada.
- KOVALEVSKY, J. (1967). *Introduction to Celestial Mechanics* Vol. 7 in "Astrophysics and Space Science Library." Translated by Express Translation Service. Springer/Reidel.
- KRARUP, T. (1973). On potential theory. Danish Geodetic Institute Report No. 6, Copenhagen, Denmark.
- LACHAPELLE, G. (1977). Physical geodesy research at the Geodetic Survey of Canada. *Proc. Symposium of the Geophysics Commission of the Pan American Institute of Geography and History*, Eds. J.G. Tanner

- and M.R. Dence. Ottawa, Canada, September, 1976. Publication of the Earth Physics Branch, Department of Energy, Mines and Resources, Ottawa, Vol. 46, No. 3, pp. 121–135.
- LACHAPELLE, G. (1978). Estimation of the geoid and deflection components in Canada. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 103–116.
- LACHAPELLE, G. (1980). Redefinition of Canadian geodetic networks. Geodetic Seminar on the Impact of Redefinition and New Technology on the Surveying Profession. Geodesy and Control Surveys Committee of the CIS, Ottawa, Canada.
- LAMBECK, K., A. CAZENAVE AND G. BALMINO (1974). Solid earth and ocean tides estimated from satellite orbit analysis. *Rev. Geophys. and Space Phys.* 12 (3), pp. 421–434.
- LAMBERT, W.D. (1930). The reduction of observed values of gravity to sea level. *Bull. Géod.*, 26, pp. 107–181.
- LAMBERT, W.D. AND F.W. DARLING (1936). Tables for determining the form of the geoid and the indirect effect on geodesy. U.S. Coast and Geodetic Survey Special Publication 199, Washington, D.C., U.S.A.
- LERCH, F.J., S.M. KLOSKO, R.E. LAUBSCHER AND C.A. WAGNER (1977). Gravity model improvement using GEOS-3 (GEM 9 and 10). Goddard Space Flight Center Report X-921-77-246, Greenbelt, U.S.A.
- LEVALLOIS, J.J. (1970). *Géodésie Générale*. Vol. III, Eyrolles.
- MACMILLAN, W.D. (1930). *The Theory of the Potential*. Dover reprint, 1958.
- MACMILLAN, W.D. (1936). *Dynamics of Rigid Bodies*. Dover reprint, 1960.
- MATHER, R.S. (1973). A solution of the geodetic boundary value problem to order  $e^3$ . Goddard Space Flight Center preprint X-592-73-11, Greenbelt, U.S.A.
- MCCULLOH, T.H. (1965). A confirmation by gravity measurements of an underground density profile based on core densities. *Geophysics* XXX (6), pp. 1108–1132.
- MERRY, C.L. (1975). Studies towards an astrogravimetric geoid for Canada. Department of Surveying Engineering Technical Report 31 University of New Brunswick, Fredericton, Canada.
- MERRY, C.L. AND P. VANÍČEK (1974a). A method for astrogravimetric geoid determination. Department of Surveying Engineering Technical Report 27, University of New Brunswick, Fredericton, Canada.
- MERRY, C.L. AND P. VANÍČEK (1974b). The geoid and datum translation components. *Canad. Surv.* 28 (2), pp. 56–62.
- MERRY, C.L. AND P. VANÍČEK (1974c). A technique for determining the geoid from a combination of astrogeodetic and gravimetric deflections. *Canad. Surv.* 28 (5), pp. 549–554.
- MOLODENSKIJ, M.S., V.F. EREMEEV AND M.I. YURKINA (1960). *Methods for Study of the External Gravitational Field and Figure of the Earth*. Translated from Russian by the Israel Program for Scientific Translations for the Office of Technical Services, U.S. Department of Commerce, Washington, D.C., U.S.A., 1962.
- MORANDO, B. (Ed.) (1970). *Dynamics of Satellites* (1969). Springer.
- MORELLI, C. (1963). The first order and absolute world gravity nets. *Boll. Geof. Teor. Appl.* 19, pp. 195–216.
- MORITZ, H. (1963). Interpolation and prediction of point gravity anomalies. Publications of the Isostatic Institute of the IAG, No. 40, Helsinki, Finland.
- MORITZ, H. (1968). On the use of the terrain correction in solving Molodenskij's problem. Department of Geodetic Science Report 108, The Ohio State University, Columbus, U.S.A.
- MORITZ, H. (1979). *Advanced Physical Geodesy*. Herbert Wichmann.
- MORRISON, F. (1972). Density layer models for the geopotential. *Am. Scientist* 60 (2), pp. 229–236.
- MUELLER, I.I. (1963). Geodesy and the torsion balance. *J. Surv. Map. Div. Proc. Am. Soc. Civ. Engg.* 89, pp. 123–155.
- MYINT-U, T. (1973). *Partial Differential Equations of Mathematical Physics*. American Elsevier.
- OFFICER, C.B. (1974). *Introduction to Theoretical Geophysics*. Springer.
- ORLIN, H. (Ed.) (1966). Extension of gravity anomalies to unsurveyed areas. American Geophysical Union Monograph 9, Washington, D.C., U.S.A.
- PICK, M., J. PÍCHA AND V. VYSKOČIL (1973). *Theory of the Earth's Gravity Field*. Elsevier.

- POLLARD, H. (1976). *Celestial Mechanics*. No. 18 in: "Carus Mathematical Monographs." Mathematical Association of America.
- RAPP, R.H. (1964). The prediction of point and mean gravity anomalies through the use of a digital computer. Department of Geodetic Science Report 43, The Ohio State University, Columbus, U.S.A.
- RAPP, R.H. (1972). Satellite orbit computations using gravity anomalies. *Studia Geoph. et Geod.* 16, pp. 1-9.
- RAPP, R.H. (1977). Potential coefficient determination from 5° terrestrial gravity data. Department of Geodetic Science Report 251, The Ohio State University, Columbus, U.S.A.
- RAPP, R.H. (1979). GEOS 3 data processing for the recovery of geoid undulations and gravity anomalies. *J. Geophys. Res.* 84 (B8), pp. 3784-3792.
- SCHWARZ, K.P. AND G. LACHAPELLE (1980). Local characteristics of the gravity anomaly. *Bull. Géod.* 54 (1), pp. 21-36.
- SEPELIN, T.O. (1974). The Department of Defense world geodetic system 1972. *Canad. Surv.* 28 (5), pp. 496-506.
- SOLLINS, A. (1947). Tables for the computation of deflections of the vertical gravity anomalies. *Bull. Géod.* 6, pp. 286-300.
- STOKES, G.G. (1849). On the variation of gravity at the surface of the earth. *Trans. Cambridge Philos. Soc.* VIII, pp. 672-695.
- STRANGE, W.E. (1982). An evaluation of orthometric height accuracy using bore hole gravimetry. *Bull. Géod.* 56(4), pp. 300-311.
- SYMON, K.R. (1971). *Mechanics*. 3rd ed., Addison-Wesley.
- TELFORD, W.M., L.P. GELDART, R.E. SHERIFF AND D.A. KEYS (1976). *Applied Geophysics*. Cambridge University Press.
- TSCHERNING, C.C. AND R. FORSBERG (1978). Prediction of deflections of the vertical. *Proc. 2nd International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, U.S. Department of Commerce, Canadian Department of Energy, Mines and Resources, Danish Geodetic Institute, Arlington, U.S.A., April. U.S. Government Printing Office, Washington, D.C., U.S.A., pp. 117-134.
- TSCHERNING, C.C. AND R.H. RAPP (1974). Closed covariance expressions for gravity anomalies, geoid undulations, and deflections of the vertical implied by anomaly degree variance models. Department of Geodetic Science Report 208, The Ohio State University, Columbus, U.S.A.
- TUCKER, R.H., A.H. COOK, H.M. IYER AND F.D. STACEY (1970). *Global Geophysics*. Elsevier.
- VALLIANT, H.D. (1971). The Canadian pendulum apparatus, design and operation. Publications of the Earth Physics Branch, Vol. 41, No. 4, Department of Energy, Mines and Resources, Ottawa, Canada.
- VANIČEK, P. AND C.L. MERRY (1973). Determination of the geoid from deflections of the vertical using a least-squares surface fitting technique. *Bull. Géod.* 109, pp. 261-279.
- VEIS, G. AND C.H. MOORE (1960). SAO differential orbit improvement program. *JPL Seminar Proceedings on Tracking, Programs, and Orbit Determination*, Jet Propulsion Laboratory, Pasadena, U.S.A., pp. 165-184.
- VENING MEINESZ, F.A. (1928). A formula expressing the deflection of the plumbline in the gravity anomalies and some formulae for the gravity field and the gravity potential outside the geoid. *Proc. Koninkl. Ned. Akad. Wetenschap* 31 (3), pp. 315-331.
- VINCENT, S., W.E. STRANGE AND J.G. MARSH (1972). A detailed gravimetric geoid of North America to Eurasia. Goddard Space Flight Center Report X-553-72-94, Greenbelt, U.S.A.
- WONG, L. AND R. GORE (1969). Accuracy of geoid heights from modified Stokes kernels. *Geophys. J. Roy. Astronom. Soc.* 18, pp. 81-91.

## PART VI

# TEMPORAL VARIATIONS

## CHAPTER 25

### CORRECTIONS FOR TEMPORAL VARIATIONS

As we have seen in Chapter 8, some temporal variations of positions and gravity field are better known than others. Thus, when it comes to correcting different geodetic quantities, be they observations, positions, or gravity field parameters, basically only two phenomena are known well enough to allow us to correct for their effects. The two phenomena are the tide and the wobble of the earth's spin axis. In both cases, the driving forces present no serious problem, and the only uncertainty is in the quantification of the earth's response.

Of the two phenomena, the tide is by far the most important. This is why the first three sections are devoted to it. In the first section, the modelling of the tidal behaviour of the actual, deformable earth in response to tidal stress—called the (earth) body tide to distinguish it from the earth's behaviour in response to sea tide—is explained. The second section enumerates and treats the individual geodetic quantities that are to be corrected. The mathematical models of the corrections are developed and discussed. The third section deals with corrections for the effects arising from sea tide. In all these sections, we rely heavily on concepts explained in §8.1 and §8.2. The last section explains corrections arising from polar motion. The motion was discussed in §5.3 and §5.4, while its effects on coordinate values and geodetic azimuth were explained in §15.2. Here we speak about the deformations caused by the wobble. It is also shown that the existing variations in the earth's spin velocity have no measurable effect on geodetic quantities.

The other phenomena discussed in Chapter 8 are not known well enough, and thus corrections for their effects are generally not considered. Vis-à-vis these phenomena, the role of geodesy changes from that of the consumer of information on temporal variations to that of the supplier of geometrical data. This second role is expanded on in the last two chapters of this part.

#### 25.1. Elastic response to tidal stress

Let us begin this section by recalling eqn. (8.6) that defines the lunar tidal potential  $W^L$ . It will be shown that all three basic tidal phenomena can be described through the tidal potential. First, the lunar tidal variation of gravity  $g$  is obtained from (8.6) simply by taking the derivative of the tidal potential along the

plumb line. Replacing the direction of the plumb line with that of the radius vector  $r$  of the point of interest and accounting for the signs, we have approximately

$$g_t^{\mathbb{C}} = -\frac{\partial W_t^{\mathbb{C}}}{\partial r} = -\frac{GM^{\mathbb{C}}}{r\rho^{\mathbb{C}}} \sum_{n=2}^{\infty} n \left( \frac{r}{\rho^{\mathbb{C}}} \right)^n P_n(\cos Z^{\mathbb{C}}). \quad (25.1)$$

An equivalent formula can easily be written for the solar contribution  $g_t^{\odot}$ .

To derive the equation for tidal variation of tilt, recall FIG. 8.3 that depicts the relation between the earth's gravity vector  $\bar{g}$ , the tidal gravity vector  $\bar{a}_t$ , and the tidal tilt  $\theta_t$  in the plane containing the two vectors. In the first approximation,  $\theta_t$  can be expressed as the ratio between the horizontal component of  $\bar{a}_t$ , and the magnitude  $g$  of the earth's gravity, the former being very much smaller. From FIG. 8.2, the horizontal component can be deduced as

$$(\bar{a}_t^{\mathbb{C}})_{\text{hor}} = \frac{1}{r} \frac{\partial W_t^{\mathbb{C}}}{\partial Z^{\mathbb{C}}}. \quad (25.2)$$

Hence the lunar contribution  $\theta_t^{\mathbb{C}}$  to tidal tilt can be written as

$$\theta_t^{\mathbb{C}} \doteq \frac{GM^{\mathbb{C}}}{rg\rho^{\mathbb{C}}} \sum_{n=2}^{\infty} \left( \frac{r}{\rho^{\mathbb{C}}} \right)^n \frac{\partial P_n(\cos Z^{\mathbb{C}})}{\partial Z^{\mathbb{C}}}. \quad (25.3)$$

A completely analogous formula is valid for the solar contribution  $\theta_t^{\odot}$ . It must be noted—see FIG. 1—that tidal tilt  $\theta_t$  is a spatial angle and should be treated as such. The two components  $\xi_t, \eta_t$  of the tilt should be considered in much the same way as the (stationary) components of the deflection of the vertical were in §6.4. We shall see later on how these components are actually evaluated.

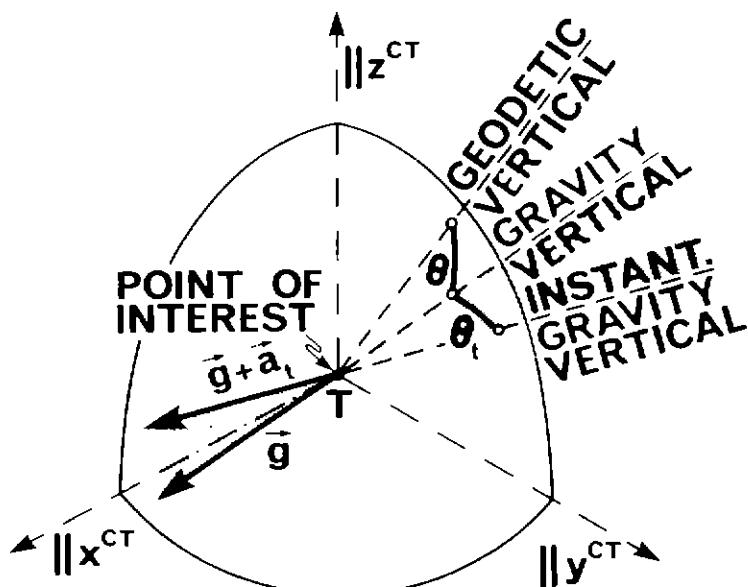


FIG. 25.1. Instantaneous deflection of the vertical.

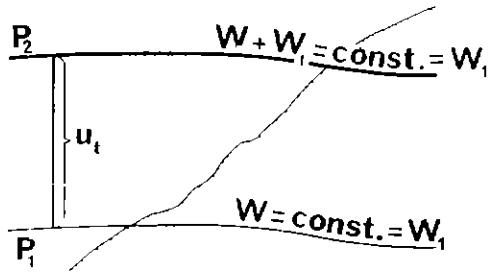


FIG. 25.2. Tidal uplift of an equipotential surface.

The tidal uplift of the earth's gravity equipotential surfaces is obtained from (6.28) by regarding  $W_t$  as a small increment  $\delta W$  to  $W$ , the sought uplift  $u_t$  as a small vertical displacement  $\delta h$  (see FIG. 2), and changing the sign due to the fact that  $W_t$  increases while  $W$  decreases. For the moon, we get

$$u_t^{\mathbb{C}} = \frac{W_t^{\mathbb{C}}}{g} = \frac{GM^{\mathbb{C}}}{gp^{\mathbb{C}}} \sum_{n=2}^{\infty} \left( \frac{r}{\rho^{\mathbb{C}}} \right)^n P_n(\cos Z^{\mathbb{C}}). \quad (25.4)$$

A parallel formula holds for the sun.

Before we turn to the geodetic implications of tidal effects, let us have a closer look at the temporal and geographical behaviour of the tidal potential. To do so, we shall confine ourselves to the predominant term in the expression for the lunar tidal potential (8.6), the second-order harmonic term:

$$W_2^{\mathbb{C}} = \frac{GM^{\mathbb{C}}}{\rho^{\mathbb{C}}} \left( \frac{r}{\rho^{\mathbb{C}}} \right)^2 P_2(\cos Z^{\mathbb{C}}) = \frac{3GM^{\mathbb{C}} r^2}{2\rho^{\mathbb{C}} C^3} \left( \cos^2 Z^{\mathbb{C}} - \frac{1}{3} \right). \quad (25.5)$$

It has been found convenient to introduce the following constant [DOODSON, 1922]:

$$D = \frac{3}{4} \frac{GM^{\mathbb{C}}}{C^3} \frac{R^2}{\rho^{\mathbb{C}}} = 2.6277 \times 10^7 \text{ cm mGal}, \quad (25.6)$$

where  $R$  is again the mean radius of the earth, and  $C^{\mathbb{C}}$  is the mean distance of the moon from the earth. The numerical value of *Doodson's tidal constant* is determined from the values of the fundamental astronomical constants adopted by the IAU in 1977 [IAU, 1977]. Using this constant, (5) can be rewritten as follows:

$$W_2^{\mathbb{C}} = 2D \left( \frac{r}{R} \right)^2 \left( \frac{C^{\mathbb{C}}}{\rho^{\mathbb{C}}} \right)^3 \left( \cos^2 Z^{\mathbb{C}} - \frac{1}{3} \right), \quad (25.7)$$

where the two ratios  $r/R$  and  $C^{\mathbb{C}}/\rho^{\mathbb{C}}$  depart from one by so small an amount that, in the first approximation, the departures may be neglected altogether.

The periodic term  $\cos^2 Z^{\mathbb{C}} - \frac{1}{3}$  can now be rewritten in terms of the observer's latitude  $\phi$ , the declination  $\delta^{\mathbb{C}}$ , and the hour angle of the moon  $h^{\mathbb{C}}$  (cf. §15.1).

Squaring (15.14) and replacing  $\Phi$  by  $\phi$ , we get

$$\cos^2 Z^\zeta = \sin^2 \phi \sin^2 \delta^\zeta + 2 \sin \phi \cos \phi \sin \delta^\zeta \cos \delta^\zeta \cos h^\zeta + \cos^2 \phi \cos^2 \delta^\zeta \cos^2 h^\zeta. \quad (25.8)$$

After some lengthy algebraic operations, we obtain

$$W_2^\zeta \doteq D [\cos^2 \phi \cos^2 \delta^\zeta \cos 2h^\zeta + \sin 2\phi \sin 2\delta^\zeta \cos h^\zeta + 3(\sin^2 \phi - \frac{1}{3})(\sin^2 \delta^\zeta - \frac{1}{3})]. \quad (25.9)$$

The declination  $\delta^\zeta$  varies with time very slowly and, therefore, the time variation of  $W_2^\zeta$  is controlled predominantly by  $h^\zeta$ . Evidently, the first term, the *sectorial contribution*  $S^\zeta$ , is responsible for the lunar semidiurnal variations; the second, the *tesseral contribution*  $T^\zeta$ , is responsible for the lunar diurnal variations; and the last, the *zonal contribution*  $Z^\zeta$ , oscillates very slowly around the constant value of the permanent tide. The reader is advised to compare this terminology with that of §20.2.

An identical expression could be developed for the solar tidal potential, with the only difference being that all the quantities would refer to the sun rather than the moon, and we would have solar semidiurnal and solar diurnal variations. Also, the Doodson constant for the sun is only about 46% of that for the moon (cf. §8.1). Putting these two potentials together, we arrive at the tidal spectrum shown in FIG. 8.4. Further, (9) shows how the tidal potential changes with latitude. These changes are plotted, for the main frequencies, in FIG. 3.

In our derivations, until now we have tacitly assumed the earth to be rigid, i.e., we have allowed neither for the earth itself to respond to the *tidal stress* nor for the effect of such an induced deformation. The question thus arises: What will the magnitude of the tidal phenomena be when observed on the surface of the real, deformable earth? As shown in the introduction to Chapter 8, deformations of a short-periodic nature are thought to behave in accordance with an elastic model. This is, therefore, how the *body tide* is usually modelled; the visco-elastic model is used only for long-periodic stresses. The elastic model shown here was first introduced by LOVE [1911]. To show how this model works, let us take the tidal uplift first, since it is the most illustrative. The ratio of the elastic radial displacement of a mass element of the real earth to the radial displacement of the corresponding element of a hypothetical fluid earth is known as the *first Love's number* and is denoted by  $h$ . Thus a fluid earth would be characterized by  $h = 1$  and a completely rigid earth by  $h = 0$ . The actual earth has Love's numbers between 0 and 1, and they depend on the degree of the spherical harmonic of the deforming force. For the degree 2, the global value of  $h$ , called  $h_2$ , has been determined under the assumption of homogeneity as 0.62 [MELCHIOR, 1978] from the semidiurnal response to the hemispherical (second harmonic) tidal stress. The diurnal response gives comparable results. These are, however, less reliable because the resonance of the earth's liquid core has more effect on the diurnal behaviour of the earth. On the other hand, for

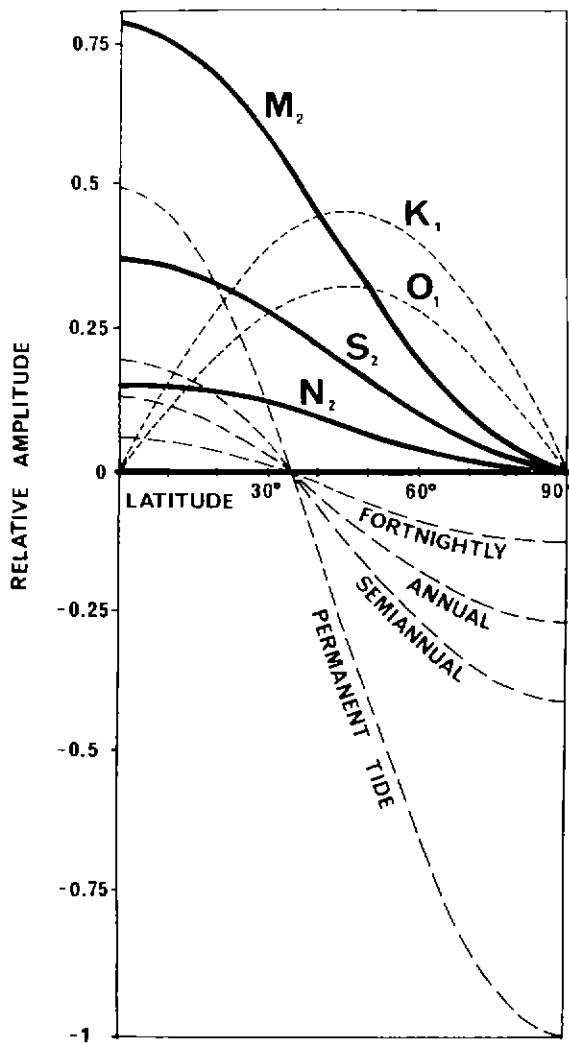


FIG. 25.3. Variation of tidal potential components with latitude.

comparison,  $h_3$ , determined from the response to the third-order harmonic of tidal stress, is equal to  $0.34 \pm 0.10$  (computed from data found in MELCHIOR [1978]). The first Love's number also varies with depth. In geodetic applications, though, only values pertaining to the earth's surface are of interest, and these are the ones we will be using here.

The ratio of the horizontal elastic displacement of an actual mass element to the horizontal displacement of the corresponding element of the hypothetical fluid earth is called *Shida's number* and is denoted by  $l$ . For this purpose, the earth is considered not only homogeneous but also isotropic (responding in the same manner in all directions) within each horizontal layer. Such an isotropic Shida's number then varies only with depth and the harmonic order of the deforming force, the same way as does the  $h$ . On the earth's surface, the value of  $l_2$  is about 0.08 [OZAWA, 1961], determined again from the analyses of tidal deformations for mostly semidiurnal frequencies.

The shape of the earth changes in response to the tidal stress; put in other words, the mass of the earth is redistributed under the action of the tidal force. The redistribution subsequently affects the earth's own gravitational field. This change,

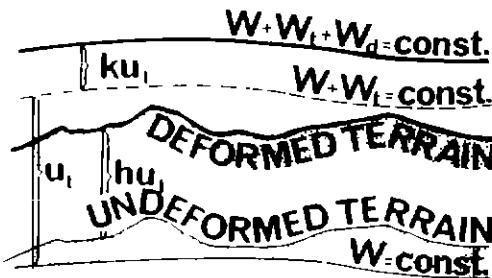


FIG. 25.4. Tidal response of deformable earth.

expressed as a change in the earth's potential, is called the potential of the deformation or, briefly, the *deformation potential*  $W_d$ . Its effect is once more most clearly seen on the case of tidal uplift; it magnifies the calculated value  $u_t$  by a factor of  $1+k > 1$ . Using (6.28) again, we can write for the uplift affected by the mass redistribution,

$$u_t(1+k) = \frac{W_t + W_d}{g}, \quad (25.10)$$

and thus, as the reader can derive,

$$k = W_d / W_t. \quad (25.11)$$

This ratio is called the *second Love's number*, and its role is seen in FIG. 4.

One can justly argue that by responding to the deformation potential, the earth itself changes its shape once more, and thus the effect of the tidal stress should change as well. This in turn will cause an additional deformation potential and so on, ad infinitum. Indeed, what we observe are the final deformations after the equilibrium has been reached. The values of all Love's numbers—and Shida's number is normally referred to as the *third Love's number*—thus reflect this equilibrium state. The value of  $k_2$  determined experimentally (predominantly again from semidiurnal frequencies) is 0.29. For comparison, the third-order harmonic deformations give  $k_3 = 0.14 \pm 0.07$  (from data found in MELCHIOR [1978]). In our applications, we shall be using only the Love numbers pertaining to second-degree harmonic deformations, and the subscript 2 will be omitted for simplicity.

## 25.2. Tidal corrections

It can readily be seen that the tidal phenomena on the surface of the real earth affect all geodetic quantities; some are affected significantly, while others are not. We have already seen how gravity, the geoidal heights, and a tilt of equipotential surfaces are influenced. We shall now show the tidal changes in and corrections to other geodetic quantities.

Let us begin with the simplest effect first: the *tidal variation of the geodetic height*  $h$  of a terrain point above a geocentric reference ellipsoid. It is evidently affected by

$$\delta r_t = \delta h_t = hu_t, \quad (25.12)$$

as can be seen from FIG. 4. Denoting  $W_2^{\text{C}} + W_2^{\text{O}}$  by  $W_2$ , the correction to the instantaneous geodetic height, to obtain a ‘mean’ geodetic height, reads

$$Oh_t \doteq -\frac{0.62}{g} W_2 \doteq -0.63 \times 10^{-6} [\text{mGal}^{-1}] W_2. \quad (25.13)$$

The *tidal variation of the geoidal height N* above a fixed geocentric ellipsoid is given by (10). Thus the correction to the instantaneous geoid, to get the ‘mean’ geoid, is

$$ON_t \doteq -\frac{1.29}{g} W_2 \doteq -1.32 \times 10^{-6} [\text{mGal}^{-1}] W_2. \quad (25.14)$$

Considering the tidal effects on the geodetic height and geoid together, we arrive at the *tidal variation of the orthometric, or normal, height H* above the geoid. It is given as

$$\delta H_t = [h - (1+k)] u_t, \quad (25.15)$$

which yields a correction to the instantaneous orthometric (normal) height:

$$OH_t \doteq 0.68 \times 10^{-6} [\text{mGal}^{-1}] W_2. \quad (25.16)$$

Naturally, the *tidal variation of the dynamically undisturbed sea level* observed from shore would be affected by the same amount  $u_t$  as the orthometric heights, except for the opposite sign. This is, however, a somewhat useless recognition since the sea level is never undisturbed. The dynamic phenomena are always present, and the problem of determining the mean sea level is tackled along the lines described in §19.1.

The derivation of the correction to instantaneous gravity is slightly more difficult. To begin with, let us suppose we are able to measure gravity at a point in space not attached to the earth. Then the *tidal variation of such absolute gravity g'* is given by the obvious equation

$$g'_t \doteq -\frac{\partial}{\partial r} (W_t + W_d). \quad (25.17)$$

Limiting ourselves to  $W_2$ , we can write

$$g'_t \doteq -\frac{\partial W_2}{\partial r} - \frac{\partial(kW_2)}{\partial r} = -(1+k) \frac{\partial W_2}{\partial r} - W_2 \frac{\partial k}{\partial r}. \quad (25.18)$$

It can be shown [LOVE, 1911] that  $k$  for  $W_2$  is approximately proportional to  $r^{-5}$  so that  $\partial k / \partial r \doteq -5k/r$ . Realizing that  $\partial W_2 / \partial r = 2W_2/r$ , we finally have

$$g'_t \doteq -\left(1 - \frac{3}{2}k\right) \frac{\partial W_2}{\partial r}. \quad (25.19)$$

On the earth’s surface ( $r \doteq R$ ), we get the correction to absolute gravity equal to

$$Og'_t \doteq \left(1 - \frac{3}{2}k\right) \frac{2W_2}{R} \doteq 0.18 \times 10^{-8} [\text{cm}^{-1}] W_2. \quad (25.20)$$

To evaluate the gravity variation observed on the earth's surface, i.e., the *tidal variation of observed gravity*  $g$ , we have to take into account the fact that the observer is also uplifted through the gravity field by an amount  $\delta r_t$ . The observed gravity change due to this displacement is equal to (cf. (12) and (6.12))

$$\delta r_t \frac{\partial g}{\partial r} \doteq -h \frac{W_2}{g} \frac{2g}{R} = -h \frac{2W_2}{R} \quad (25.21)$$

Thus the total tidal effect is

$$g_t \doteq -\left(1 + h - \frac{3}{2}k\right) \frac{2W_2}{R}, \quad (25.22)$$

and the corresponding correction reads

$$Og_t \doteq 0.37 \times 10^{-8} [\text{cm}^{-1}] W_2. \quad (25.23)$$

The investigation of tidal effects on distances and angles requires a special mathematical apparatus. Let us first denote the displacement vector of a mass element within the earth by  $\bar{v} = (v_x, v_y, v_z)^T$ . Then the symmetrical matrix that describes differential deformations in Cartesian coordinates,

$$\begin{aligned} \epsilon &= \frac{1}{2} \left[ \nabla \bar{v}^T + (\nabla \bar{v}^T)^T \right] \\ &= \begin{bmatrix} \frac{\partial v_x}{\partial x} & \frac{1}{2} \left( \frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x} \right) & \frac{1}{2} \left( \frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x} \right) \\ \frac{1}{2} \left( \frac{\partial v_y}{\partial x} + \frac{\partial v_x}{\partial y} \right) & \frac{\partial v_y}{\partial y} & \frac{1}{2} \left( \frac{\partial v_y}{\partial z} + \frac{\partial v_z}{\partial y} \right) \\ \frac{1}{2} \left( \frac{\partial v_z}{\partial x} + \frac{\partial v_x}{\partial z} \right) & \frac{1}{2} \left( \frac{\partial v_z}{\partial y} + \frac{\partial v_y}{\partial z} \right) & \frac{\partial v_z}{\partial z} \end{bmatrix}, \end{aligned} \quad (25.24)$$

is called the *strain tensor* [CONDON AND ODISHAW, 1967]. When dealing with the earth's deformations, the Cartesian coordinate system is not the most convenient. It is far more preferable to work with strain in spherical or, even better, in geodetic coordinates. In geodetic coordinates, the strain tensor reads (cf. LOVE [1927] after replacing the second spherical coordinate by  $\frac{1}{2}\pi - \phi$ ) approximately

$$\epsilon \doteq \begin{bmatrix} \frac{\partial v_r}{\partial r} & \frac{\partial v_\phi}{\partial r} - \frac{v_\phi}{r} + \frac{\partial v_r}{r \partial \phi} \\ \frac{\partial v_\phi}{\partial r} - \frac{v_\phi}{r} + \frac{\partial v_r}{r \partial \phi} & \frac{\partial v_\phi}{r \partial \phi} + \frac{v_r}{r} \\ \frac{1}{r \cos \phi} \frac{\partial v_r}{\partial \lambda} + \frac{\partial v_\lambda}{\partial r} - \frac{v_\lambda}{r} & \frac{\partial v_\lambda}{r \partial \phi} - \tan \phi \frac{v_\lambda}{r} + \frac{1}{r \cos \phi} \frac{\partial v_\phi}{\partial \lambda} \end{bmatrix}$$

$$\begin{aligned}
 & \left[ \begin{array}{c} \frac{1}{r \cos \phi} \frac{\partial v_r}{\partial \lambda} + \frac{\partial v_\lambda}{\partial r} - \frac{v_\lambda}{r} \\ \frac{\partial v_\lambda}{r \partial \phi} - \tan \phi \frac{v_\lambda}{r} + \frac{1}{r \cos \phi} \frac{\partial v_\phi}{\partial \lambda} \\ \frac{1}{r \cos \phi} \frac{\partial v_\lambda}{\partial \lambda} + \tan \phi \frac{v_\phi}{r} + \frac{v_r}{r} \end{array} \right] \\
 = & \begin{bmatrix} e_{rr} & e_{r\phi} & e_{r\lambda} \\ e_{\phi r} & e_{\phi\phi} & e_{\phi\lambda} \\ e_{\lambda r} & e_{\lambda\phi} & e_{\lambda\lambda} \end{bmatrix}. \tag{25.25}
 \end{aligned}$$

If we are interested only in the earth's surface, it makes sense to talk only about the *surface*, or horizontal, *strain*, expressed by a two-dimensional tensor consisting of the southeastern four elements,

$$\epsilon' = \begin{bmatrix} e_{\phi\phi} & e_{\phi\lambda} \\ e_{\lambda\phi} & e_{\lambda\lambda} \end{bmatrix}. \tag{25.26}$$

Now, in geodetic coordinates, the components of the *tidal displacement vector*  $\bar{v}$  are [MELCHIOR, 1978]

$$v_r = \delta r_t = hu_t = h \frac{W_t}{g}, \quad v_\phi = \frac{l}{g} \frac{\partial W_t}{\partial \phi}, \quad v_\lambda = \frac{l}{g} \frac{\partial W_t}{\cos \phi \partial \lambda}, \tag{25.27}$$

where the derivative with respect to  $\lambda$  is carried out through the hour angle, which is the only function of  $\lambda$ . Substituting these into (26), we get, after some rearrangement,

$$\epsilon' \doteq \frac{h}{Rg} W_t \mathbf{I} + \frac{l}{Rg} \mathbf{D}^2(W_t), \tag{25.28}$$

where  $\mathbf{D}^2$  is the following symmetrical matrix differential operator

$$\mathbf{D}^2 = \begin{bmatrix} \frac{\partial^2}{\partial \phi^2} & \frac{2}{\cos \phi} \frac{\partial^2}{\partial \phi \partial \lambda} \\ \frac{2}{\cos \phi} \frac{\partial^2}{\partial \phi \partial \lambda} & \frac{1}{\cos^2 \phi} \frac{\partial^2}{\partial \lambda^2} + \tan \phi \frac{\partial}{\partial \phi} \end{bmatrix}, \tag{25.29}$$

and  $\mathbf{I}$  is the unit matrix. Numerical evaluation gives

$$\epsilon' = 1.0 \times 10^{-15} [\text{cm}^{-1} \text{mGal}^{-1}] W_t \mathbf{I} + 1.3 \times 10^{-16} [\text{cm}^{-1} \text{mGal}^{-1}] \mathbf{D}^2(W_t). \tag{25.30}$$

In these expressions,  $e_{\phi\phi}$  is merely a relative deformation (extension or compression) in the direction of the meridian. Similarly,  $e_{\lambda\lambda}$  is a relative deformation in the prime vertical. Thus the relative deformation  $e_\alpha$  in the direction of azimuth  $\alpha$  is given by

the expression (see (16.50))

$$e_\alpha = \mathbf{q}^T(\alpha) \tilde{\boldsymbol{\epsilon}} \mathbf{q}(\alpha), \quad (25.31)$$

where

$$\tilde{\boldsymbol{\epsilon}} = \begin{bmatrix} e_{\phi\phi} & \frac{1}{2}e_{\phi\lambda} \\ \frac{1}{2}e_{\lambda\phi} & e_{\lambda\lambda} \end{bmatrix}, \quad \mathbf{q}(\alpha) = \begin{bmatrix} \cos \alpha \\ \sin \alpha \end{bmatrix}. \quad (25.32)$$

Evaluation of the first and second derivatives of  $W_t$  leads to expressions more complicated than the evaluation of the radial derivative. It can be seen from (9) that these derivatives have a different form for each frequency. Thus *tidal horizontal strain* has a different character for different frequencies, and (28) cannot be further simplified. For instance, the deformation in the meridian is given as

$$e_0 = e_{\phi\phi} \doteq 1.0 \times 10^{-15} [\text{cm}^{-1} \text{mGal}^{-1}] W_t + 1.3 \times 10^{-16} [\text{cm}^{-1} \text{mGal}^{-1}] \frac{\partial^2 W_t}{\partial \phi^2}, \quad (25.33)$$

and for the  $M_2$  frequency, one gets

$$e_0^{(M_2)} \doteq 2.6 \times 10^{-8} S^{\mathbb{C}} + 0.34 \times 10^{-8} \frac{\partial^2 S^{\mathbb{C}}}{\partial \phi^2} = (2.6 \times 10^{-8} \cos^2 \phi - 0.7 \times 10^{-8} \cos 2\phi) \cos^2 \delta^{\mathbb{C}} \cos 2h^{\mathbb{C}}. \quad (25.34)$$

The pattern of horizontal strain in all directions  $\alpha$  for the  $M_2$  frequency is shown in

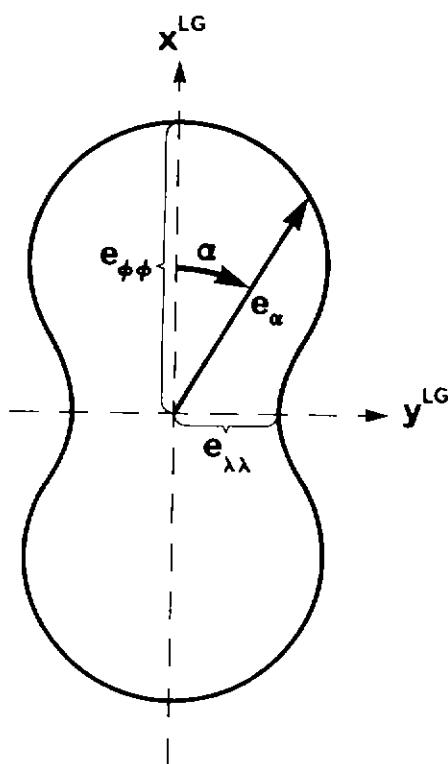


FIG. 25.5. Horizontal strain pattern of  $M_2$  frequency for latitude  $\phi = 45^\circ$ .

FIG. 5. This pattern is the pedal curve (cf. FIG. 16.16) to the ellipse with axes  $e_{\phi\phi}$  and  $e_{\lambda\lambda}$ , and it represents the graphical form of (31). There are other ways to depict horizontal strain, such as strain ellipses, as we shall see in §27.4.

All the strain effects, and thus the *tidal variations of horizontal distances*  $S$  as well, are of the order of a few parts in  $10^{-8}$  at most. Hence tidal corrections to distances have to be applied only to very precise measurements such as those obtained through radio-interferometry (cf. §16.1). The correction to a horizontal distance can be written as

$$OS_t = -\mathbf{q}^T(\alpha) \sum_{\kappa} \tilde{\epsilon}_{\kappa} \mathbf{q}(\alpha) S, \quad (25.35)$$

where  $\tilde{\epsilon}_{\kappa}$  are surface strain tensors (cf. (32)) evaluated for all desired tidal frequencies  $\kappa$ . The derivation of the formulae for the elements of the surface strain tensor for sectorial, tesseral, and zonal constituents is left to the reader. *Tidal variations of horizontal angles*  $\omega$  are far below the noise level of the measuring techniques.

The next geodetic quantity affected by tidal stress is the astronomical deflection of the vertical. The *tidal variation of the astronomical deflection of the vertical*  $\theta'$  (cf. FIG. 21.3) can be visualized as being composed of two constituents:

- (a) the tilt of the equipotential surface that is obviously equal to  $(1+k)\theta_t$ , and
- (b) the shift of the terrain point.

While the equipotential surface changes from  $W = \text{const.}$  to  $W + W_t + W_d = \text{const.}$ , the observing station  $P$  on the earth's surface is displaced to  $P'$  due to horizontal displacement  $Rl\theta_t$  (see FIG. 6). Hence, the total effect at the earth's surface, in the plane containing the tide-inducing celestial body, is (cf. (3))

$$\theta'_t = (1+k-l)\theta_t = \frac{1+k-l}{gR} \frac{\partial W_t}{\partial Z}. \quad (25.36)$$

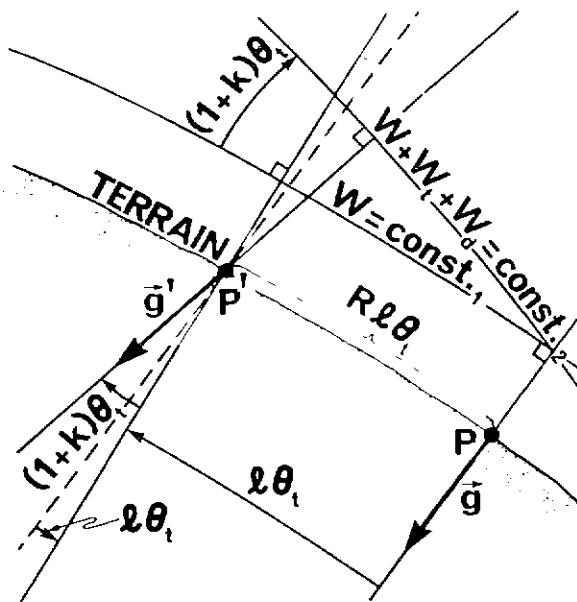


FIG. 25.6. Tidal effect on the astronomical deflection of the vertical.

Taking into account the sign convention for deflections (21.18), the corrections to the astrodeflection components in the meridian and prime vertical are

$$\begin{aligned} O\xi'_t &\doteq -1.94 \times 10^{-15} [\text{cm}^{-1} \text{mGal}^{-1}] \frac{\partial W_2}{\partial \phi}, \\ O\eta'_t &\doteq -1.94 \times 10^{-15} [\text{cm}^{-1} \text{mGal}^{-1}] \frac{\partial W_2}{\cos \phi \partial \lambda}. \end{aligned} \quad (25.37)$$

The tilt of the terrain measured with respect to an equipotential surface is again affected by the distortion of the equipotential surfaces and that of the earth's surface. The effect of the first distortion is our old friend the tilt  $(1+k)\theta_t$ . The second is the tilt given by the following formula:

$$\frac{\partial}{\partial S} (hu_t) \doteq \frac{h}{g} \frac{\partial W_t}{\partial S} = h\theta_t, \quad (25.38)$$

where the derivative is carried out in the direction in which the tilt is measured. The dependence on Shida's number that we experienced with astrodeflections disappears here because of the relative nature of tilt. The total *tidal variation of observed terrain tilt*  $\theta''$  is thus given by  $(1+k-h)\theta_t$ . The corresponding correction is then

$$O\theta''_t = - \frac{(1+k-h)}{g} \frac{\partial W_t}{\partial S} \doteq -0.68 \times 10^{-6} [\text{mGal}^{-1}] \frac{\partial W_t}{\partial S}, \quad (25.39)$$

where the derivative is taken with respect to distance  $S$ . Clearly, it is convenient to express tilt in azimuth  $\alpha$  as a function of tilt in the meridian and prime vertical directions. This can be done by means of (16.80), and we shall do it in the context of the next geodetic quantity.

The *tidal variation of a levelled height difference*  $\delta l$  can be evaluated from the terrain tilt using simple geometry. The situation is shown in FIG. 7. Using (16.80)

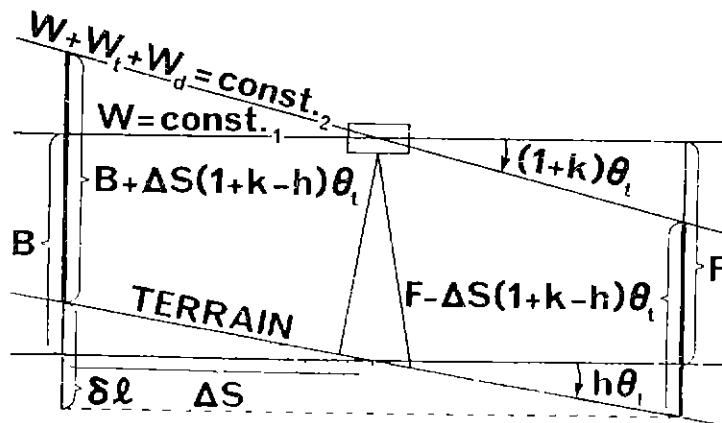


FIG. 25.7. Tidal effect on levelled height difference

and (39), we get the appropriate correction as [VANIČEK, 1980]

$$\begin{aligned} O\delta l_t &= \frac{1+k-h}{gR} 2\Delta S \left( \cos \alpha \frac{\partial W_2}{\partial \phi} + \sin \alpha \frac{\partial W_2}{\cos \phi \partial \lambda} \right) \\ &\doteq 2.14 \times 10^{-15} [\text{cm}^{-1} \text{mGal}^{-1}] \Delta S \left( \cos \alpha \frac{\partial W_2}{\partial \phi} + \sin \alpha \frac{\partial W_2}{\cos \phi \partial \lambda} \right). \end{aligned} \quad (25.40)$$

It should be noted that the *tidal variation of a vertical angle*  $\nu$  is composed of the variations in the deflection of the vertical and the variation in the tilt of the earth's surface. The derivation of the appropriate correction is left to the reader. Let us only state here that under present circumstances, the application of such a correction does not seem warranted, because the observational accuracy is several orders of magnitude lower than even the maximum value of the correction.

TABLE 1 summarizes the maximum possible ranges of all the above variations for the moon and the sun. It should be kept in mind that all these ranges depend on latitude and some on azimuth as well. Therefore, the listed maxima do not apply universally.

To calculate any of the above corrections, it is necessary to evaluate numerically either the tidal potential or some of its horizontal derivatives. This has to be done for every point of interest and for the time of observation, and eqn. (9) can be used for the task. For the evaluation, we have to know the declination and the hour angle of both the moon and the sun for the instant of time we are interested in. The concept of one possible technique, useful when routine evaluation of these corrections is contemplated, will be explained in the next section in the context of sea tide effects. Also, it should be pointed out that, in practice, adequate accuracy can be achieved

TABLE 25.1  
Maximum possible range of tidal variations experienced on the earth's surface (disregarding permanent tide)

Geodetic quantity	Symbol	#	Eqn.		Range	Remarks
			Lunar	Solar		
Geodetic height	$h$	13	33 cm	15 cm		Also equal to the range in $r$ .
Orthometric (or normal) height	$H$	16	36 cm	17 cm		Also equal to the range in relative uplift of undisturbed sea level/equipotential surface—see TABLE 8.2.
Geoidal height	$N$	14	69 cm	32 cm		Also equal to the range in absolute uplift of undisturbed sea level.
Absolute gravity	$g'$	20	95 $\mu\text{Gal}$	44 $\mu\text{Gal}$		Gravity variation at a point in space considered here.
Observed gravity	$g$	23	194 $\mu\text{Gal}$	90 $\mu\text{Gal}$		Also see TABLE 8.2.
Horizontal distance	$S$	35	$8.0 \times 10^{-8} S$	$3.7 \times 10^{-8} S$		Depends on azimuth.
Astrodeflection	$\theta'$	37	0.021''	0.010''		Depends on azimuth.
Observed terrain tilt	$\theta''$	39	0.012''	0.005''		Also see TABLE 8.2. Depends on azimuth.
Levelled height difference	$\delta l$	40	0.056 mm	0.026 mm		Per one km of levelling. Depends on azimuth.

by including just the five predominant frequencies  $\{M_2, S_2, N_2, O_1, K_1\}$ . There are, however, more appropriate algorithms available that generate the tidal potential without any frequency discrimination.

It should be reiterated that throughout this section, we have assumed the earth's response to be laterally homogeneous and isotropic. In reality, there may be non-negligible deviations from homogeneity and isotropy present due to local topography and horizontal variations in the earth's elastic properties; the latter are caused by variations in geological structure. These deviations affect tilt and horizontal strain more seriously than gravity; for further details see, e.g., BAKER AND LENNON [1976].

### 25.3. Corrections due to sea tide effects

In §8.2, the effect of the load of tidal water on the lithosphere was mentioned. It turns out that loading is not the only effect tidal water has on solid-earth phenomena. There are two more effects that have to be taken into account when temporal variations of geodetic quantities are investigated: the *gravitational attraction of tidal water*, and the gravitational effect of the deformation caused by the load, called the *indirect effect of tidal water*. Obviously, the latter effect is akin to the effect of the deformation potential discussed in the previous two sections. All three effects of tidal water are usually treated together much the same way as the various manifestations of the body tide were treated in the last two sections.

The easiest effect to evaluate is the gravitational attraction. It requires only the knowledge of tidal water distribution, i.e., the cotidal and corange charts (cf. §8.1), and the position of the point of interest with respect to these water masses. Evidently, the earth's rheology is irrelevant in the evaluation of this effect. As we shall see, this is why this effect serves as a 'reference effect' for the other two, in a fashion similar to the response of the fluid earth in the body-tide theory.

To show how the three phenomena are treated, let us begin once more with vertical displacements. We shall denote the vertical displacement of the earth's surface under the load by  $u_l$ , the displacement of a gravity equipotential surface due to the attraction of water by  $u_a$ , and the displacement of an equipotential surface due to the indirect effect by  $u_i$ . Then the elastic model of the response of the earth and its gravity field to tidal water can be characterized by a system of functions similar to Love's numbers. Introduced by MUNK AND MACDONALD [1960], these are called *load numbers* and are denoted by  $h'$ ,  $k'$ , and  $l'$ .

The load numbers are once more defined as ratios; namely,

$$h' = u_l/u_a, \quad k' = u_i/u_a, \quad (25.41)$$

and, similarly,

$$l' = v_l/v_a,$$

where  $v$  denotes a horizontal displacement in the direction of the load. Load numbers, like Love's numbers, are functions of both depth and the horizontal

dimension of the load, i.e., the degree of the spherical harmonic function that can be used to describe the load. In our applications, we shall deal only with the surface response and thus only with surface load numbers. Also, conforming to custom, we shall denote the dependence of load numbers on wavelength by subscript  $n$ . Thus, the elastic response to *point load*, meaning a load with a negligibly small horizontal dimension, is characterized by  $h'_\infty, k'_\infty, l'_\infty$ . All three effects are intertwined and cannot be readily separated. As in the case of the Love numbers, the values of load numbers will reflect an equilibrium situation.

At this stage it is advantageous to introduce the gravitational (attraction) potential  $W_w$  of the tidal waters, which we shall call simply the *sea tide potential*. Denoting the amplitude of the sea tide (at point  $\vec{r}$  and the moment  $\tau$  of interest) by  $Z$ , we can write the following expression for the potential at  $A$  (cf. (22.30)):

$$W_w(\vec{r}_A, \tau) = -G \iint_{\mathcal{G}} \frac{Z(\vec{r}, \tau) \sigma_w(\vec{r}, \tau)}{\rho(\vec{r}_A, \vec{r})} d\mathcal{G}, \quad (25.42)$$

where  $Z$  is equal to zero on land,  $\rho$  is once more the chord distance, and the integration is carried out over the entire earth's surface  $\mathcal{G}$ . Considering the density of water  $\sigma_w$  to be constant throughout the seas, a good enough assumption for this purpose, we can rewrite (42) as (valid for the instant  $\tau$  for which  $Z$  is taken)

$$W_w(\phi_A, \lambda_A) \doteq -G \sigma_w R^2 \iint_{\mathcal{G}} \frac{Z(\phi, \lambda)}{\rho(\phi_A, \lambda_A, \phi, \lambda)} d\nu, \quad (25.43)$$

where  $d\nu$  is again the solid angle element, and  $R$  is the mean radius of the earth.

Applying the same technique as we used to derive (8.4) for  $W_t$  (or (20.48) for  $W_g$ ) and using spherical approximation, we can rewrite (43) in terms of an integral of the infinite series of Legendre's polynomials in  $\cos \psi$ , where  $\psi$  is, as before, the spherical angle between the point of interest  $(\phi_A, \lambda_A)$  and the dummy point  $(\phi, \lambda)$ :

$$W_w(\phi_A, \lambda_A) \doteq -G \sigma_w R \iint_{\mathcal{G}} Z(\phi, \lambda) \sum_{n=0}^{\infty} P_n(\cos \psi) d\nu. \quad (25.44)$$

Interchanging the summation with the integration, we have

$$W_w(\phi_A, \lambda_A) \doteq - \sum_{n=0}^{\infty} G \sigma_w R \iint_{\mathcal{G}} Z(\phi, \lambda) P_n(\cos \psi) d\nu. \quad (25.45)$$

Using the Legendre decomposition formula (20.51) for  $P_n(\cos \psi)$ , we finally get the sea tide potential as a series of spherical harmonics:

$$W_w(\phi_A, \lambda_A) \doteq \sum_{n=0}^{\infty} \sum_{m=0}^n (W_w)_{nm}, \quad (25.46)$$

where the potential coefficients are given as (cf. (20.52))

$$\begin{aligned} \left\{ \begin{array}{l} A_{nm}(W_w) \\ B_{nm}(W_w) \end{array} \right\} &\doteq -2G \sigma_w R \frac{(n-m)!}{(n+m)!} \\ &\times \iint_{\mathcal{G}} Z(\phi, \lambda) \left\{ \begin{array}{l} \cos m\lambda \\ \sin m\lambda \end{array} \right\} P_{nm}(\sin \phi) d\nu. \end{aligned} \quad (25.47)$$

Here, the zonal coefficients ( $m = 0$ ) have to be multiplied by one-half. Equation (44) may be equivalently rewritten in a closed form,

$$W_w(\phi_A, \lambda_A) \doteq -\frac{G\sigma_w R}{\sqrt{2}} \iint_{S_1} Z(\phi, \lambda) (1 - \cos \psi)^{-1/2} d\nu, \quad (25.48)$$

when we realize (cf. (20.49) for  $r = r_A = 1$ ) that

$$\sum_{n=0}^{\infty} P_n(\cos \psi) = \frac{1}{\sqrt{2}} (1 - \cos \psi)^{-1/2}. \quad (25.49)$$

As the  $W_w$  is an attraction potential, it is easy to see (cf. (4)) that

$$u_a = W_w/g. \quad (25.50)$$

Evidently, the vertical displacement due to the gravitational attraction of tidal water can be evaluated either through a series of spherical harmonics, using (46), or, by means of (48), through the following convolution integral:

$$u_a(\phi_A, \lambda_A, \tau) = \iint_{S_1} K_a(\phi_A, \lambda_A, \phi, \lambda) Z(\phi, \lambda, \tau) d\nu, \quad (25.51)$$

where the integration kernel, homogeneous and isotropic, is—see FIG. 8—

$$K_a = -\frac{G\sigma_w R}{\sqrt{2} g} (1 - \cos \psi)^{-1/2} \doteq -0.0315 (1 - \cos \psi)^{-1/2}. \quad (25.52)$$

Thus, given the distribution of water within the seas at the desired instant  $\tau$ , we can perform the above integration and obtain the displacement due to attraction at the desired point  $(\phi_A, \lambda_A)$ . As an illustration, a 3000 km by 3000 km patch of sea with a

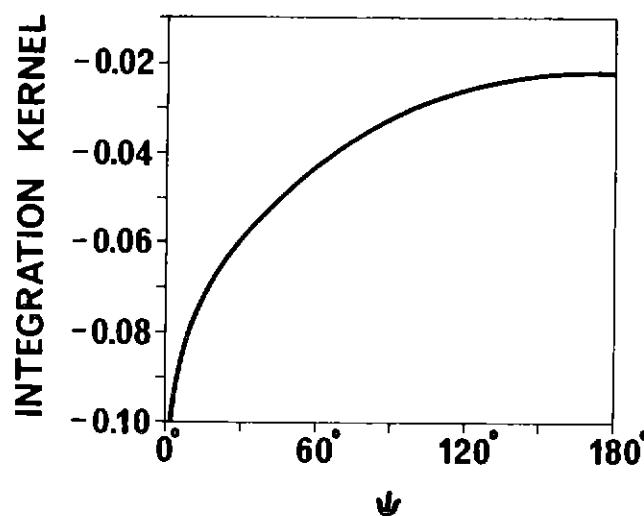


FIG. 25.8. Integration kernel  $K_a$  for gravitational attraction.

one metre-high tide at a distance of  $\psi = \pi/2$  has a gravitational effect which causes the uplift of an equipotential surface of  $-0.7$  centimetres.

It should be easy by now to see that the other displacements can be written as (see (41))

$$\boxed{u_1 = \frac{1}{g} \sum_{n=0}^{\infty} h'_n (W_w)_n,}$$

$$u_1 = \frac{1}{g} \sum_{n=0}^{\infty} k'_n (W_w)_n \quad \text{and} \quad v_1 = \frac{1}{g} \sum_{n=0}^{\infty} l'_n \frac{\partial (W_w)_n}{\partial \psi}. \quad (25.53)$$

There are not enough observations available to determine the earth's response, and thus the load numbers, experimentally. Instead, the load numbers are obtained from the rheological models of the earth. FARRELL's [1972] results calculated from the standard Guttenberg–Bullen 'A' rheological model (see, e.g., ALTERMAN ET AL. [1961]) of the earth are shown in FIG. 9.

An alternative form of (53) can be derived by interchanging the summation with the integration, as we have done for  $W_w$ . Then we can write, for instance, for  $u_1$

$$\boxed{u_1(\phi_A, \lambda_A, \tau) = \iint_K K_1(\phi_A, \lambda_A, \phi, \lambda) Z(\phi, \lambda, \tau) d\nu,} \quad (25.54)$$

where the integration kernel, again homogeneous and isotropic, is equal to

$$\boxed{K_1 = -\frac{G\sigma_w R}{g} \sum_{n=0}^{\infty} h'_n P_n(\cos \psi).} \quad (25.55)$$

The reader is advised to derive the corresponding equations for  $u_1$  and  $v_1$  as an exercise. We observe, as we did in §22.1, that these integration kernels are also

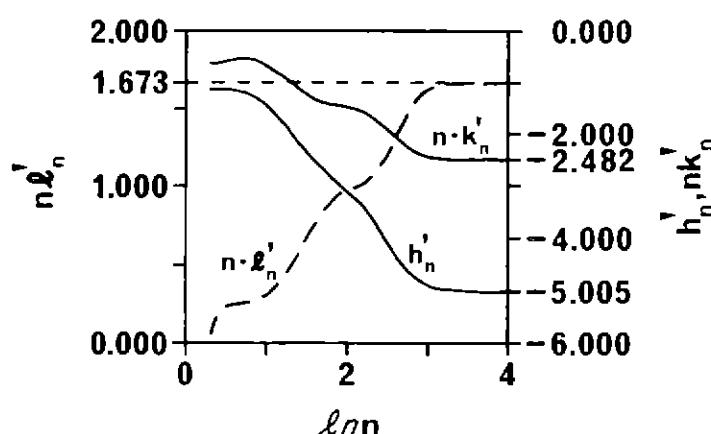


FIG. 25.9. Load numbers.

Green's functions, viewed from the point of view of the boundary value problems; they are often referred to as such in the literature.

The widest range of various Green's functions for elastic displacements, tilt, strain, and gravity response to surface loads was calculated by FARRELL [1972], based on LONGMAN's [1962; 1963] foundations. Farrell used three alternative earth models to generate the Green functions: the Guttenberg-Bullen 'A' model (for Green's function of vertical displacement, see FIG. 10), and two of HARKRIDER's [1970] variants of the same model. The latter two differ from the former only by the top 1000 km which are replaced by oceanic or continental shield structures respectively. The results show that except for the nearest 500 km from the load, the difference in the models, and thus the difference in the load numbers, does not play any role. Investigations by other researchers, e.g., BEAUMONT AND LAMBERT [1972] and ZSCHAU [1976], lead to the same conclusion: For more distant loads, the three models are equivalent to a few percent.

For shorter distances, Green's functions can no longer be regarded as homogeneous; they vary with local geology by as much as a few tens of percents. Other problems arise in modelling the earth's response in regions immediately adjacent to shore, i.e., very close to the load. There, not only local geology but also the shape of the terrain play significant roles. In addition, the accuracy of existing cotidal charts, or sea tide models, in coastal areas is not as high as one would like. Little is known at present about the possible anisotropy of the actual earth's response (and thus of more realistic Green's functions). All things considered, the sea tide effects may be evaluated at present to only a few tens of percents. This is, of course, better than nothing when we realize that the effects can be quite significant: sea tide effect on tilt near the coast for instance, may be up to an order of magnitude larger than the body tide tilt [LENNON AND VANÍČEK, 1970].

Because load numbers are defined in much the same way as Love's numbers, the effects of sea tide on various geodetic quantities are given by equations parallel to those in §25.2, where, naturally,  $W_l$  is replaced by  $W_w$ . The corrections are expressed in terms of the same combinations of load numbers as the tidal corrections are of

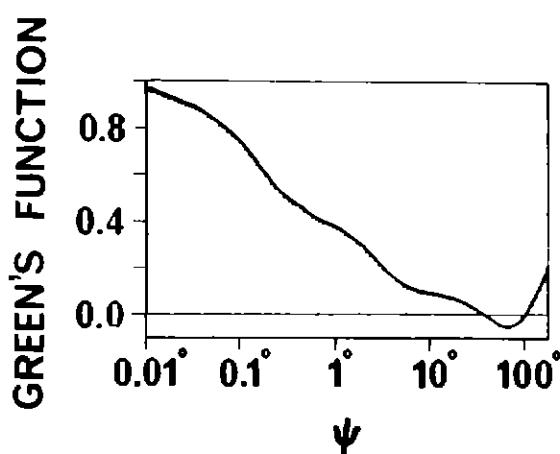


FIG. 25.10. Green's function  $K_1$  normalized by  $G\sigma_w R / (gh_\infty)$ , expressed as a function of distance from load.

Love's numbers. The main difference is that the equations have to be applied to all wave numbers  $n$  separately and then summed up. Alternatively, the convolution integrals can be used. There is nothing to stop us from using a combination of both—a truncated series for distant areas and the convolution integral over the residual sea tide in the immediate neighbourhood (spherical cap) of the point of interest—in the same manner in which the combined geoid was evaluated in §24.4. A more economical alternative based on a preintegration of Green's functions has been proposed by GOAD [1979].

The necessity of re-evaluating the desired effects for different time instants can be overcome by expressing the sea tide  $Z$  as a complex function of position  $(\phi, \lambda)$ . For each frequency  $\kappa$  (corresponding to  $M_2$ ,  $S_2$ ,  $O_1$ , etc.) the tidal magnitude  $Z$  at an instant  $\tau$  can be written as

$$Z_\kappa(\phi, \lambda, \tau) = A_\kappa(\phi, \lambda) \cos[\kappa\tau - \beta_\kappa(\phi, \lambda)], \quad (25.56)$$

where the amplitude  $A_\kappa$  and phase lag  $\beta_\kappa$  are obtained from cotidal and corange charts as functions of position alone. Denoting

$$Z_\kappa^*(\phi, \lambda) = A_\kappa(\phi, \lambda) \exp(-i\beta_\kappa(\phi, \lambda)), \quad i = \sqrt{-1}, \quad (25.57)$$

we can rewrite (56) as (cf. §3.1)

$$Z_\kappa(\phi, \lambda, \tau) = \operatorname{re}[Z_\kappa^*(\phi, \lambda) \exp(i\kappa\tau)]. \quad (25.58)$$

Thus, knowing  $Z_\kappa^*$ , we can evaluate the tidal magnitude  $Z_\kappa$  for any instant  $\tau$  merely by multiplying  $Z_\kappa^*$  by a complex function of time, namely,  $\exp(i\kappa\tau)$ . (Note that the same technique may indeed be used for calculating the tidal potential  $W_t$  in the tidal corrections in §25.2.)

Let us now denote any of the effects  $e$  of sea tide by  $f^e(\phi, \lambda, \tau)$ . Then we can write a formula, similar to (58),

$$f^e(\phi, \lambda, \tau) = \operatorname{re} \sum_\kappa f_\kappa^{e*}(\phi, \lambda) \exp(i\kappa\tau), \quad (25.59)$$

and all we need to evaluate the total effect  $e$ , for a point in time, are the functions  $f_\kappa^{e*}(\phi, \lambda)$  for all the considered frequencies. These complex functions of position are calculated from the convolution integrals of the usual kind:

$$f_\kappa^{e*}(\phi, \lambda) = \iint_{\mathcal{G}} K_e(\phi, \lambda, \phi', \lambda') Z_\kappa^*(\phi', \lambda') d\nu, \quad (25.60)$$

where  $K_e$  is the appropriate Green's function for the effect  $e$ . The convolution integrals are evaluated numerically using one of any number of existing schemes (see, e.g., BOWER [1970], GOAD [1979]), or are replaced either fully or partially by the appropriate truncated series of spherical harmonics, as we saw earlier.

Even the direct evaluation of the  $f_\kappa^{e*}$  for each point may be bypassed; the spatial distribution of these complex effects may be approximated by a linear form with an

arbitrarily chosen base  $\phi$ :

$$f_{\kappa}^{e*}(\phi, \lambda) = \tilde{\Phi}^T(\phi, \lambda) c_{\kappa}^{e*}, \quad (25.61)$$

much the same as was done with the geoid in §22.4. The only difference is that this time the coefficients are complex. By doing this, only the vectors of complex coefficients  $c_{\kappa}^{e*}$  have to be stored: they are generally different for each frequency  $\kappa$  and each effect  $e$ . The coefficients can be determined through least-squares regression or some other technique (cf. §22.4). Once the coefficients are known, then any desired correction can be determined easily from the obvious formula

$$Oe(\phi, \lambda, \tau) = -f^e(\phi, \lambda, \tau) = -\operatorname{re} \sum_{\kappa} \exp^*(i\kappa\tau) \tilde{\Phi}^T(\phi, \lambda) c_{\kappa}^{e*}, \quad i = \sqrt{-1}.$$

(25.62)

As we have seen in the previous sections, not all the corrections to geodetic quantities are independent. Therefore, under certain circumstances, it may be advisable to store the information, such as the coefficients shown above, for only the minimum set of independent effects. These independent effects may be selected in a variety of ways. For instance, denoting the vertical displacement of an equipotential surface by  $u_g$ , i.e.,

$$u_g = u_a + u_i, \quad (25.63)$$

the most natural set of independent effects is  $(u_1, u_g, v_1, g_w)$ . These in turn depend on the following linear combinations of load numbers:  $h'_n$ ,  $1 + k'_n$ ,  $l'_n$ , and  $1 + (2/n)h'_n - ((n+1)/n)k'_n$  respectively (cf. (13), (14), (25), and (22)). All other effects, and thus all other corrections, can be derived from this minimal set. For example, the correction to relative vertical displacement of the earth's surface (equivalent to the *correction to the instantaneous orthometric height* —cf. (16)) for tidal frequency  $\kappa$  is

$$(OH_w)_{\kappa} = -(u_1)_{\kappa} + (u_g)_{\kappa}. \quad (25.64)$$

Similarly, the *correction to the levelled height difference* in direction  $\alpha$  and frequency  $\kappa$  is (cf. (40))

$$(O\delta l_w)_{\kappa} = \frac{2\Delta S}{gR} \left( \cos \alpha \frac{\partial(OH_w)_{\kappa}}{\partial \phi} + \sin \alpha \frac{\partial(OH_w)_{\kappa}}{\cos \phi \partial \lambda} \right). \quad (25.65)$$

The development of the rest of the formulae is left to the reader.

Finally, if the main, say five, tidal frequencies  $\kappa \equiv \{M_2, S_2, N_2, O_1, K_1\}$  are considered, then only twenty complex vectors  $c_{\kappa}^{e*}$  have to be stored to generate any desired correction for any point at any time. The sum of the partial contributions for the above five frequencies should approximate the sea tide well enough for any geodetic purpose.

## 25.4. Corrections due to polar motion deformations, and other causes

The kinematical aspects of polar motion were explained in §5.3 and §5.4, and the polar motion deformations were mentioned in §8.4; however, we have not yet shown how the motion produces any deformations. To describe the deformation undergone by the earth and its gravity field in response to polar wobble, let us first recall the formula for the centrifugal potential  $W_c$  (6.26). It can be rewritten, as a function of time, in a handier form: namely,

$$W_c(\tau) = \frac{1}{2}\omega^2 r^2 \cos^2\phi(\tau). \quad (25.66)$$

This formula, of course, assumes a rigid earth; we shall see later how to correct for this simplification.

The latitude in (66) may be regarded as the mean latitude  $\phi$  (in the CT system of coordinates (cf. §15.1)) plus the change  $\delta\phi(\tau)$  due to polar motion. Then (66) can be rewritten as

$$W_c(\tau) \doteq \frac{1}{2}\omega^2 r^2 (\cos^2\phi - \sin 2\phi \ \delta\phi(\tau)). \quad (25.67)$$

Clearly, the first term is merely the stationary part of the centrifugal potential; the second term may be called the *polar motion potential*  $W_p$ :

$$W_p(\tau) \doteq -\frac{1}{2}\omega^2 r^2 \sin 2\phi \ \delta\phi(\tau). \quad (25.68)$$

The latitude change  $\delta\phi$  can be simply expressed as a function of the observed time-varying coordinates of the instantaneous pole ((15.7) and (15.63)). Its peak-to-peak amplitude is about  $0.5''$  (cf. FIG. 5.8), and it has two basic periods: the Chandler and the annual. The equipotential surfaces of the earth's gravity field, as they are affected by  $W_p$ , are shown in FIG. 11. A two-dimensional expression for  $W_p$  (taken as a function of both  $\phi$  and  $\lambda$ ) used, for instance, by LAMBECK [1980], may be regarded as an alternative to the above eqn. (68).

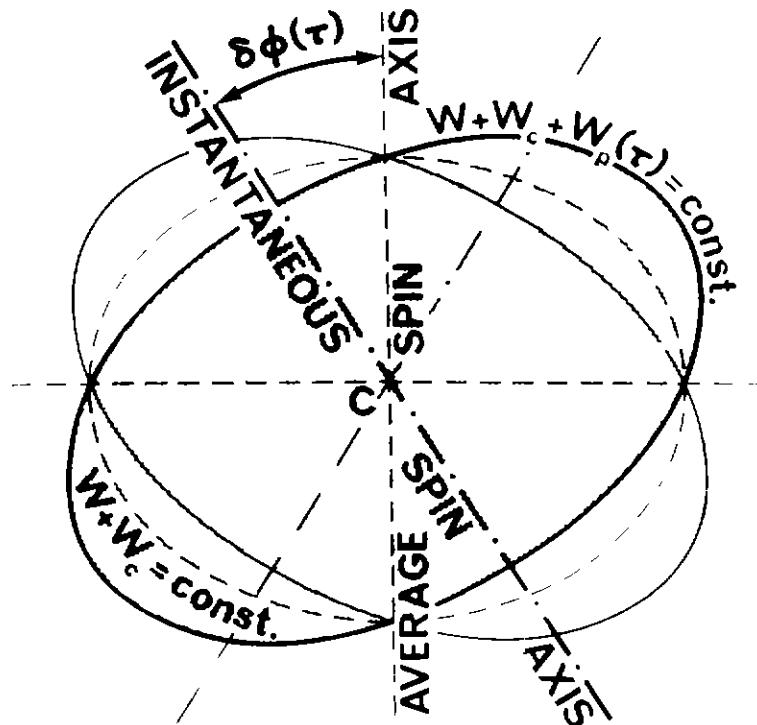
Considering the earth still as rigid, we can follow the same approach as we did in §25.1 with tidal effects, and derive the equations for the three basic polar motion effects of interest to geodesists. Taking the *polar motion uplift* of equipotential surfaces,  $u_p$ , first, we obtain

$$u_p = \frac{W_p}{g} \doteq -\frac{\omega^2 R^2}{2g} \sin 2\phi \ \delta\phi. \quad (25.69)$$

Substitution for  $\omega$ ,  $R$ ,  $g$  yields

$$u_p \doteq -11.0[\text{km}] \sin 2\phi \ \delta\phi. \quad (25.70)$$

Note that the uplift on the equator and at the poles is zero. For  $\phi = 45^\circ$  and for the range of  $0.5''$  in  $\delta\phi$ , we get the maximum range in  $u_p$  of 2.7 cm, which is only a few percent of  $u_t$ .

FIG. 25.11. Polar motion potential  $W_p$ .

The *polar motion gravity variation* on the surface of the rigid earth is given by

$$g_p \doteq -\frac{\partial W_p}{\partial r} = \omega^2 R \sin 2\phi \delta\phi. \quad (25.71)$$

Substituting for  $\omega$  and  $R$ , we have

$$g_p \doteq 3.4[\text{Gal}] \sin 2\phi \delta\phi. \quad (25.72)$$

Here again, the changes along the equator and on the poles are zero. The maximum range occurs at latitude  $\phi = 45^\circ$  and equals about  $8.2 \mu\text{Gal}$  for the maximum range of  $\delta\phi$ , again only a few percent of the tidal variation.

The last effect is the *polar motion variation of tilt* of equipotential surfaces. For the rigid earth, we get

$$\theta_p = \frac{(\delta g_p)_{\text{hor}}}{g} = \frac{\partial W_p}{gR\partial\phi} \doteq -\frac{\omega^2 R}{g} \cos 2\phi \delta\phi. \quad (25.73)$$

After substitution, we have

$$\theta_p = -0.0034 \cos 2\phi \delta\phi. \quad (25.74)$$

This time there is no tilt at latitude  $45^\circ$ , and the maximum range of  $0.0017''$  is reached at the poles and the equator. It is not worth the effort to even calculate the *polar motion horizontal strain*. Since the tidal horizontal strain is already extremely

small, the polar motion induced strain is surely negligible in any geodetic application.

To evaluate the effects of polar motion on the real, deformable earth, we have to take the earth's rheology into consideration. If the elastic response were a good enough approximation, then evidently the Love numbers  $h_2$ ,  $k_2$ , and  $l_2$  would describe the situation; note that the wave number of the deforming force is 2, the same as the predominant part of the tidal force. Then all the formulae for tidal corrections developed in §25.2 could be used, without any qualification, just by substituting  $W_p$  for  $W_t$  (or for  $W_2$ ). The mechanics of this will be left to the reader.

The visco-elastic response to *polar motion stress* is not known. Only one polar wobble effect has been actually observed and for this only on one frequency, Chandler's. This is the sea level variation (see, e.g., HOLLAND AND MURTY [1970], CURRIE [1975]). Taking the amplitude of the Chandler component of  $\delta\phi$  to be about  $0.2''$ , the observed sea level variation on the elastic earth's surface should be  $(1 + k_2 - h_2)u_p$ , which gives an amplitude of about  $0.7[\text{cm}]\sin 2\phi$ . VANÍČEK [1978] determined a mean amplitude of  $0.97 \pm 0.30$  cm from actual sea level data along the coasts of the United States corresponding to an average latitude of  $\phi = 45^\circ$ . The variation of gravity with Chandler's frequency and amplitude of  $2.7 \mu\text{Gal}$  reported by GOODKIND [1978] is, in that author's opinion, somewhat questionable.

Clearly, the polar motion effects are very small and thus of no consequence to routine geodetic work. Nevertheless, in very precise surveys of a global nature based on observations to extraterrestrial objects, the effect can be felt. In the future, when positioning and gravity measuring systems become more accurate, corrections due to polar motion deformations will be applied as a matter of routine. Note that  $k_2 g_p$  is the gravitational effect on satellites; only the attraction effect of the deformation bulge is felt by satellites since they are not earthbound.

Do the variations in the earth's rotation rate have an appreciable effect on the earth's deformations? To answer this question let us again consider the centrifugal potential  $W_c$ . When the changes in  $a$  and  $\phi$  in response to the earth's spin velocity variations  $\delta\omega(\tau)$  are neglected, then the change of  $W_c$  in response to  $\delta\omega(\tau)$  is

$$W_r(\tau) \doteq 2W_c \frac{\delta\omega(\tau)}{\omega}. \quad (25.75)$$

This quantity may be called the *potential due to spin velocity variations*. Then, taking the most rapid change in the earth's rotation,  $10 \mu\text{s}$  per day (cf. §5.4), we arrive at the upper limit for the relative change:

$$\left| \frac{\delta\omega(\tau)}{\omega} \right| < 1.2 \times 10^{-10}. \quad (25.76)$$

Thus, considering the magnitude of  $W_c$  (see §6.1) and even allowing for the changes in  $a$  and  $\phi$  to add significantly to the above  $W_r$ , the total effect will still be negligible. In fact, the *varying earth's spin velocity induced deformations* are not measurable even with the most precise instrumentation, and so far they can safely be forgotten in geodetic work.

Various geodetic quantities are, of course, also affected by the other phenomena enumerated in Chapter 8. None of these phenomena, however, is as well known, and thus as modellable (predictable) as the ones dealt with in this chapter. Thus, even when the effects may be more significant than those treated in this chapter, their prediction is also more questionable. Any decision as to whether to correct or not must be based on an assessment of whether it is more harmful to correct using poorly known corrections or not to correct at all. We do not believe a generally valid prescription can be given here, and knowledgeable geophysists should be consulted before any such decision is reached.

## CHAPTER 26

### DETECTION OF VERTICAL MOVEMENTS

The kinds of movements that are to be detected and quantified by geodetic methods were described in §8.2 to §8.4. These movements may be either continuous (e.g., postglacial rebound) or discontinuous (e.g., the creeping motion of two adjacent tectonic plates) in space. They may also be either continuous (e.g., movements due to the extraction of water from the ground) or discontinuous—episodic—(e.g., coseismic movements) in time. In addition, it is sometimes helpful to divide the sought movements into those linear in time (e.g., movements in response to sediment loading) and those accelerated (e.g., movements preceding earthquakes—precursors).

There are also different kinds of geodetic data that contain information on the movements. These data may be either discrete or continuous in time; they may be either absolute or relative; they may pertain to a point, a line, or an area. The data, which may be of vastly different accuracies, may be collected either for the specific purpose of detecting and quantifying the movements or for an altogether different purpose.

The above is indicative of the main problem here: One must put together the various types of available data with the models designed to detect and describe the various kinds of movements. The first section discusses the existing types of geodetic data, their availability and accuracy. The second section is devoted to the study of the relation between the time variations of gravity on the one hand, and time variations of heights and vertical displacements on the other hand. It also specifies the reference system in which the vertical displacements are evaluated. The third section deals with models used for displacement profiles, while the last section tackles the areal approach to modelling.

#### 26.1. Sources of information on vertical movements

As stated above, from the temporal point of view, the geodetic data may be either continuous or discrete in time, the former being preferable but more expensive to get. Similarly, from the spatial point of view, areal coverage is the best but also the most expensive. As an overwhelmingly valid rule, there is never enough data around, and one is well advised to use, in the process of detecting vertical crustal movements, all the available data even though they may have originally been collected for completely different purposes.

There are four fundamental kinds of data that can be considered suitable for vertical movement determination: (a) sea level variations, (b) repeated vertical positions, (c) tilt changes, and (d) gravity variations.

(a) The most direct type of data are the *sea level variations* as recorded by automatic tide gauges; they are recorded basically for oceanographical purposes. We have already had a close look at these records within the context of vertical datum determination (§19.1); now we shall have another look at them from the point of view of vertical crustal movement detection.

We have seen that the noise in the records due to the water level variations can be filtered out to a certain extent, if auxiliary data with which to construct the linear filter are available. From the crustal movement point of view, the signal is composed of the linear term  $c_E(\tau - \tau_0)$ —after subtracting the eustatic water rise—and the episodic movements hidden in the residuals (cf. (19.1)). Clearly, the linear term corrected for eustatic rise should reflect the linear vertical movement of the land, with respect to the mean sea level, at the locality of the tide gauge. If we consider for the moment that the eustatic correction is perfectly known, then the linear movement can be determined with an accuracy of about  $\sigma = 2$  cm per century. This holds true under the condition that the sea level record for at least 30 years' duration and at least the main kinds of auxiliary data (cf. §19.1) are available [VANÍČEK, 1978]. The residuals (cf. FIG. 19.3) are significantly contaminated by the unmodelled effects, i.e., by the crudeness of the earlier described linear model. In spite of this fact, sea level records may supply valuable information on accelerated movements; the signal (i.e., the episodic movements) of four or more months' duration with an amplitude of at least 10 cm should be visible on the noisy background.

(b) The next kind of data are *repeated vertical positions*. It should not be difficult to see that if a point has moved vertically between two determinations of its vertical position, this shows directly as a change of position. Conversely, if a change in vertical position is experienced, then it is symptomatic of a vertical movement. So far, no system capable of continuously measuring variations in vertical positions exists; only positioning that is discrete in time is practicable. As we have seen in §15.3, the existing absolute, vertical point positioning techniques have sufficient accuracy to be useful only in studies of the most pronounced vertical movements, such as sizeable coseismic displacements. The same also holds true about most of the terrestrial, relative positioning techniques (cf. §16.1 and §16.4).

(c) Only results from levelling of higher orders appear to be of general value in the context of this chapter. The most elementary piece of useful data is the elevation difference, denoted here as  $\Delta l_{ij}$ , along a segment connecting two bench marks, levelled at two different epochs  $\tau_1$  and  $\tau_2$ . The interpretation of such a *relevelled segment* is shown in FIG. 1; clearly, a relevelled segment must be viewed as yielding information, discrete in time, on tilt changes between the two bench marks involved. These segments may be levelled either for crustal movement studies or simply for vertical position determination, usually the latter. Hence, their distribution in space and time is usually not optimal for the study of movements. The spatial configuration may be scattered, linear, or areal; the time sampling may be either random or confined to some sharply defined narrow time intervals. All these characteristics

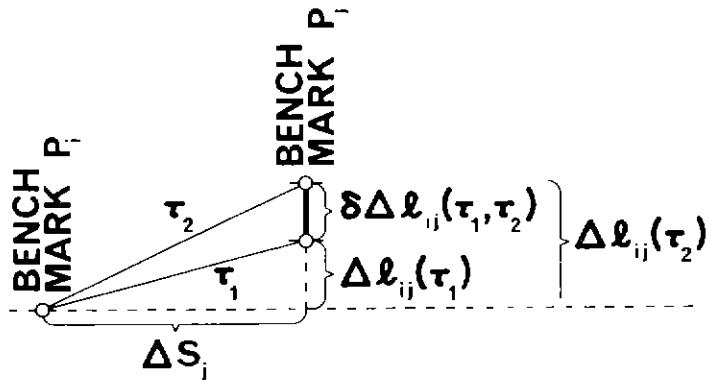


FIG. 26.1. Segment relevelled once.

have to be taken into account when the appropriate mathematical model for the movements is designed.

There are several important points pertaining to levelling data that should be discussed here. As we have seen in §16.4, to obtain from observed elevation differences any kind of proper height differences, one has to account for the effect of the actual gravity field. This is done in terms of an additive correction to the levelled height difference (cf. (16.93) and (19.5)). Therefore, the value of the height-difference difference

$$\delta \Delta H_{ij}(\tau_1, \tau_2) = \Delta H_{ij}(\tau_2) - \Delta H_{ij}(\tau_1), \quad (26.1)$$

in any height system is affected not only by the relative vertical motion of  $P_j$  with respect to  $P_i$  and by systematic and random errors arising from the levelling process, but also by the difference in the two corrections (to  $\Delta l_{ij}(\tau_1)$  and to  $\Delta l_{ij}(\tau_2)$ ) for the actual gravity effect. One way to practically eliminate the effect of gravity (for a complete discussion, see §26.2) and to significantly reduce some of the other errors, is to require that the two levellings between  $P_i$  and  $P_j$  follow the same route and have comparable sight lengths. When these requirements are satisfied, errors related to geographical location (e.g., the settling of rod and instrument supports), to height and height gradient (e.g., the residual refraction—see §19.2), and to the direction of levelling (e.g., possible solar radiation effect) are significantly suppressed. Even with these requirements in effect, there still may be some troublesome errors present, and the interpretation of the height-difference differences should be done with caution and, preferably, in conjunction with some geophysical evidence (see, e.g., CHI ET AL. [1980]). More will be said about these in §26.3.

There exist other data on tilt changes: the most obvious are *lake level variations* as registered by water level gauges. These instruments, identical with tide gauges used in recording sea level variations, continually collect data for hydrological purposes. The lake level records cannot be used in the same way as the sea level data because the lake level fluctuation, from year to year as well as within each year, is considerably more pronounced than that of the sea level. This is mainly due to precipitation and man's actions. The recorded data are thus usable only in a differenced fashion: pairing together gauges on the same lake, the difference of the

two records represents the difference in the vertical displacements of the two gauges. As in the case of sea level, the records should be rid of as much of the noise originating in water dynamics as possible. This requirement, however, is not nearly as important as in the case of sea level, because much of the noise is common to the two records and is thus eliminated through the differencing. The evaluation of the difference of two linear in time vertical displacements is shown conceptually in FIG. 2. By dividing the observed difference by the distance of the two gauges, the linear tilt variation is obtained, referred to the local equipotential surface. For an illustration of the usage of lake level data in crustal movement studies, the reader is referred to the publication of the COORDINATING COMMITTEE ON GREAT LAKES BASIC HYDRAULIC AND HYDROLOGIC DATA [1977].

*Point tilt variations* can be observed with tiltmeters. These instruments, based on a variety of physical principles, measure the variations in mutual tilt of the bedrock they are mounted on and the local equipotential surface. To be of general use, the tiltmeters have to be sufficiently sensitive; the usual requirement is for the sensitivity to be at least 1 ms of arc, i.e., 5 nanoradians. At this level of sensitivity, it is always dubious whether the recorded long periodic and secular tilt is real or spurious. It may, and often does, reflect the behaviour of the instrument itself, or that of the instrument's micro-environment, and is not representative of the region of interest [CABANISS, 1978; BRAGARD, 1980]. A somewhat more hopeful approach to tilt measurements appears to be in the use of longer base hydrostatic tiltmeters—see, e.g., BOWER [1973]. These instruments measure and record the tilt of two points several metres to several hundreds of metres apart. The stability of the instrument is better, and the recorded tilt can more readily be regarded as being representative of the whole surrounding area.

(d) The last kind of data on vertical movements are *repeated gravity observations*. Variations in gravity are caused either by a vertical displacement of the point of

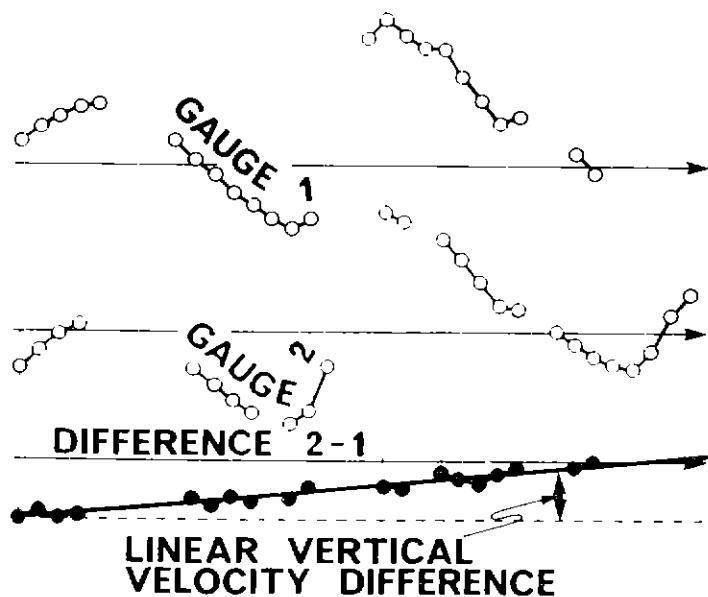


FIG. 26.2. Lake level variations.

observation or by a redistribution of masses within the earth. The question of the interrelation of gravity and height variations will be discussed more fully in the next section; let it suffice here to say that gravity changes are usually indicative of vertical movements. Hence, being the cheapest available sort of data of its kind [LAMBERT AND VANÍČEK, 1979], it is usually wise to employ gravity variations for detection purposes. Repeated gravity observations, discrete in time and space, are abundant from projects having nothing to do with vertical movements. In addition, precise gravimeters are sometimes used to observe specially designed local gravity networks [LAMBERT AND BEAUMONT, 1977]. Continually recording gravimeters, accurate to a fraction of a microgal, also exist. Some of them possess even a good, long term stability, e.g.,  $3 \mu\text{Gal}$  per year [GOODKIND, 1978].

## 26.2. Interdependence of temporal variations of gravity and heights

In §8.2, it was shown how much gravity may change as a result of postglacial rebound. Tidal variations of gravity were discussed in §8.1 and §25.2, variations due to sea tide loading in §25.3, and polar wobble induced fluctuations in §25.4. As we know, there are other effects that cause the gravity to change: plate tectonics, local subsidence, water table fluctuations, etc. Except for the coseismic variations, the most spectacular changes are those accompanying preseismic movements in the stress-accumulation regions along active faults. These may reach a hundred microgals within a space of one year [BEAUMONT, 1976]. Gravity changes caused by local subsidence display the same rates as those associated with postglacial rebound, i.e., around ten microgals per year [BEAUMONT, 1976]. Underground water table fluctuations vary considerably in magnitude from point to point, as do the gravity changes caused by these fluctuations. According to LAMBERT AND BEAUMONT [1977], these changes, which are predominantly of a seasonal character, may reach some tens of microgals. Finally, the intraplate tectonic movements of the lithosphere are responsible for variations of the order of hundredths of a microgal per year [BEAUMONT, 1976], which cannot be measured with present instrumentation. The situation is shown in FIG. 3.

The main problem with observed gravity variations is that of interpretation: How much of the observed change is due to mass redistribution within the earth, and how much is due to actual *vertical displacement*? To help with deciphering the gravity message, JACHENS [1978] derived the bounds for the ratios of gravity to height changes corresponding to different physical phenomena. His results, arrived at through numerical modelling (simulation), are shown in FIG. 4. For comparison, the ratio obtained for the body tide is added to the figure; it is calculated for the observed gravity (cf. (25.23)) and the geodetic height changes (cf. (25.13)). Using this diagram, observed gravity changes can be translated to vertical displacements, if the physical cause of the movement is known. It should be noted that neither the effect of water table fluctuations nor that of volcanic filling of cavities, and similar phenomena, can be shown on the diagram; the corresponding ratios go to infinity because there are no height changes ( $\delta h = 0$ ) arising from these phenomena.

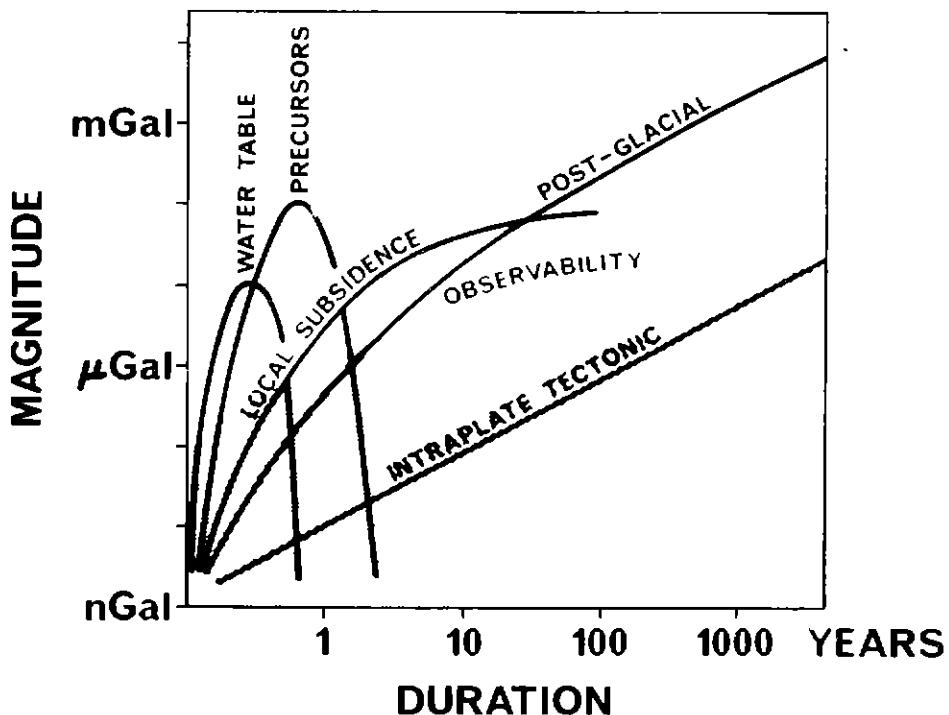


FIG. 26.3. Characteristic long term variations of gravity.

Evidently, there is little that can be meaningfully inferred from observed gravity changes alone if the cause of the movement is not known, except that a vertical displacement may have occurred. On the other hand, observed elevation changes may be taken as fairly representative of the actual vertical displacement; the influence of gravity variations on the observed elevation differences is relatively small. To prove this, let us first show that a change in the observed elevation difference  $\Delta l$  is, for all practical purposes, equivalent to the change in orthometric height difference  $\Delta H^O$ , even when the gravity variation is quite extreme.

(a) Recalling (1), the change in the orthometric height difference of two bench marks  $P_i, P_j$  can be written as (cf. (16.92) and (19.5))

$$\delta \Delta H_{ij}^O = \delta \Delta l_{ij} + \delta DC_{ij} + \delta OC_{ij}, \quad (26.2)$$

where  $\delta DC_{ij}$  is the change in the dynamic correction, and  $\delta OC_{ij}$  is the change in the orthometric correction. If the releveling has followed the same route as the original levelling, this change can be due only to a difference  $\delta g$  to gravity and the difference in heights. Differentiation of the dynamic plus orthometric corrections and substitution for the mean gravity along the plumb line yield [VANÍČEK ET AL., 1980]

$$\begin{aligned} \delta DC_{ij} + \delta OC_{ij} &\doteq \overline{\delta g}_{ij} \frac{\Delta l_{ij}}{\bar{g}} + \left( \delta g_i + \delta \nabla g_i \frac{H_i}{2} + \frac{\nabla g_i}{2} \delta H_i \right) \frac{H_i}{\bar{g}} \\ &\quad - \left( \delta g_j + \delta \nabla g_j \frac{H_j}{2} + \frac{\nabla g_j}{2} \delta H_j \right) \frac{H_j}{\bar{g}}, \end{aligned} \quad (26.3)$$

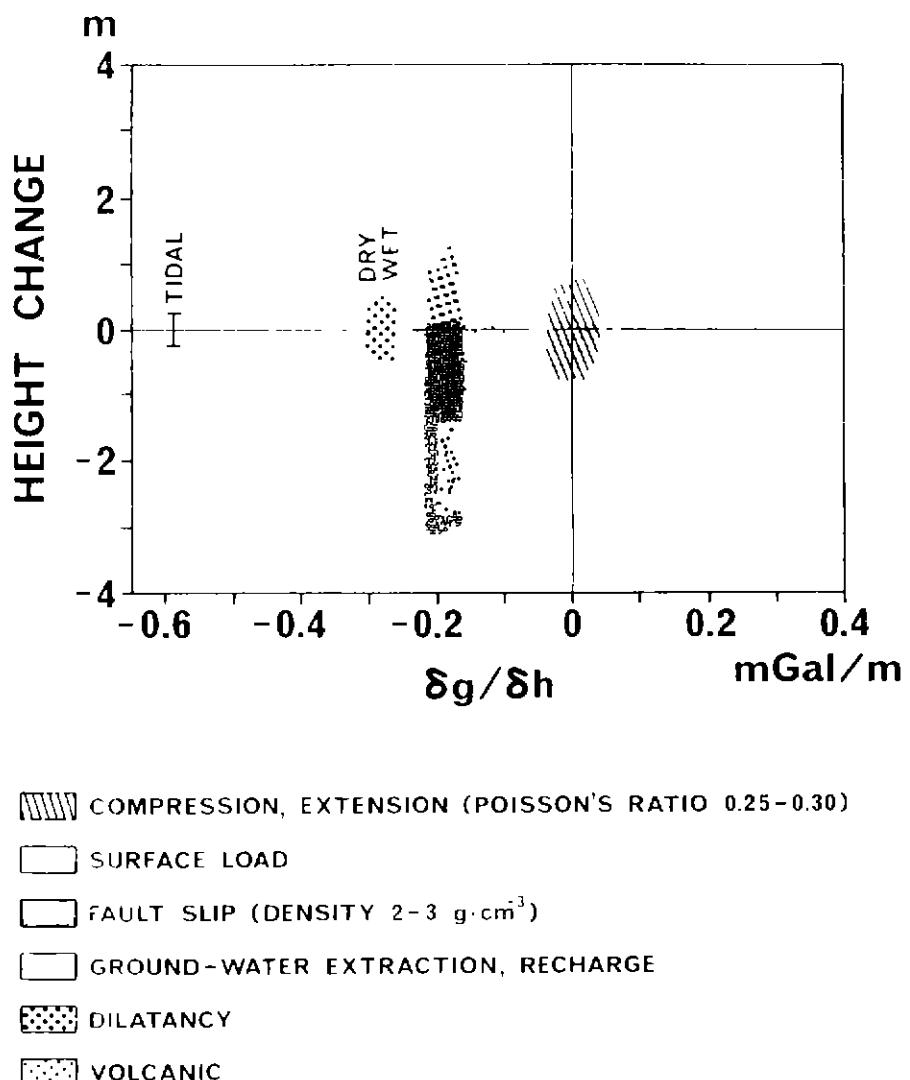


FIG. 26.4. Ratios of gravity to height changes for different physical phenomena. (Courtesy of Dr. R. C. JACHENS [1979], U.S. Geological Survey.)

where  $\bar{\delta}g_{ij}$  is the average surface gravity change between  $P_i$  and  $P_j$ ,  $\nabla g$  is the vertical gravity gradient, and  $\bar{g}$  is some global mean gravity. The reader can satisfy himself that, even under extreme conditions,  $\delta DC_{ij} + \delta OC_{ij}$  is of the order of 1 mm at most.

(b) So far nothing has been said about the effect of the gravity change on the geoid. The aforementioned orthometric height difference  $\Delta H_{ij}^O$  at any epoch  $\tau$  must be considered to be related to an *instantaneous geoid*. By how much the instantaneous geoid departs from the mean geoid is the question which we want to address presently. To this end, let us take Stokes's formula (22.17) that gives the geoidal height  $N$  above a geocentric reference ellipsoid as a function of, say, free air gravity anomalies  $\Delta g$ . Denoting the mean anomaly in a ring at spherical distance  $\psi$  from the point of interest by  $\bar{\Delta}g$ , we have

$$\bar{\Delta}g(\psi) = \frac{1}{2\pi} \int_0^{2\pi} \Delta g(\psi, \alpha) d\alpha. \quad (26.4)$$

The use of the  $F$  function (see (22.72)) gives

$$N \doteq \frac{R}{2\bar{g}} \int_0^\pi \overline{\Delta g}(\psi) F(\psi) d\psi. \quad (26.5)$$

Then the change in  $N$  due to the gravity change is

$$\delta N \doteq \frac{R}{2\bar{g}} \int_0^\pi \delta \overline{\Delta g}(\psi) F(\psi) d\psi, \quad (26.6)$$

where  $\delta \overline{\Delta g}(\psi)$  is the mean change of the gravity anomaly within the spherical distance  $\psi$ .

Let the gravity and height changes occur only within the radius  $\psi_{\max}$  of the point of interest, i.e.,  $\delta \overline{\Delta g} = 0$  for  $\psi > \psi_{\max}$ . Then integration of (6) by parts yields [VANÍČEK ET AL., 1980]

$$\delta N \doteq -\frac{R}{2\bar{g}} \int_0^{\psi_{\max}} \frac{\partial}{\partial \psi} \delta \overline{\Delta g}(\psi) \chi(\psi) d\psi, \quad (26.7)$$

where  $\chi$  was defined by (22.74). For  $\psi < 10^\circ$ ,  $\chi(\psi)$  may be approximated by  $2.3\psi$  [LAMBERT AND DARLING, 1936]. Moreover, considering the free air anomaly, we get

$$\delta \overline{\Delta g}(\psi) \doteq \overline{\delta g}(\psi) + 0.31[\text{mGal m}^{-1}] \overline{\delta H}(\psi), \quad (26.8)$$

where  $\overline{\delta g}(\psi)$  is the average surface gravity change, and  $\overline{\delta H}(\psi)$  is the average vertical displacement, both at spherical distance  $\psi$ . Substituting this result back into (7), we finally obtain

$$\delta N \doteq -7.4[\text{mGal}^{-1} \text{m}] \int_0^{\psi_{\max}} \psi \frac{\partial \overline{\delta g}(\psi)}{\partial \psi} d\psi - 2.3 \int_0^{\psi_{\max}} \psi \frac{\partial \overline{\delta H}(\psi)}{\partial \psi} d\psi.$$

$$(26.9)$$

For illustration, let us assume a conical model for  $\overline{\delta g}$  and  $\overline{\delta H}$ , where the maxima  $\delta g_{\max}, \delta H_{\max}$  are reached for the point of interest ( $\psi=0$ ), and both  $\overline{\delta g}$  and  $\overline{\delta H}$  decrease linearly to zero at  $\psi = \psi_{\max} \leq 10^\circ$  (FIG. 5). Then the above formula reduces to

$$\delta N \doteq (7.4[\text{mGal}^{-1} \text{m}] \delta g_{\max} + 2.3 \delta H_{\max}) \psi_{\max}. \quad (26.10)$$

Clearly, for smaller conical features of the order of 100 km, the geoidal change is small: for  $\delta g_{\max} = 100 \mu\text{Gal}$ ,  $\psi_{\max} = 1^\circ$ , we get  $\delta N = 1.3 \text{ cm} + 0.04 \delta H_{\max}$ . Models of other shapes give even smaller values of  $\delta N$ .

The last point that should be made here concerns the temporal variations of the geoid resulting from its definition through the mean sea level. As pointed out already in §19.1, the consequence of the standard definition is that both the shape and the potential, and thus even the size, of the geoid, change with the eustatic water rise thus reducing all the heights around the world by 0.6 mm to 1.0 mm every year.

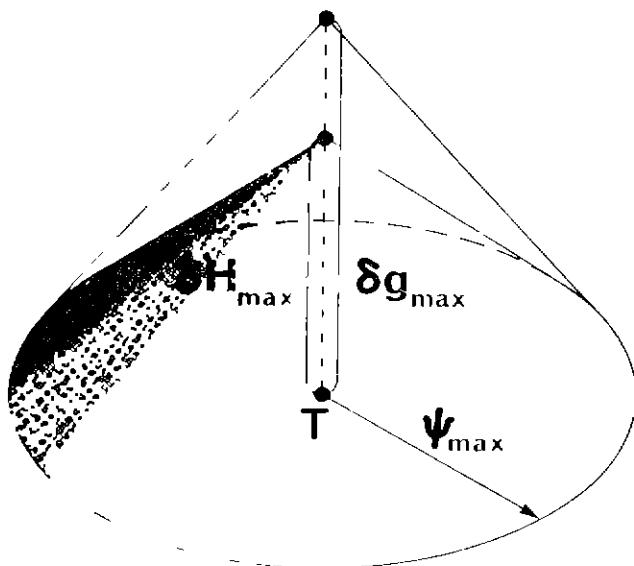


FIG. 26.5 Conical model for vertical displacement and gravity changes.

Clearly, even if the geoidal potential is fixed, the shape would still change in response to the redistribution of water and ice on the earth. It is convenient in terms of vertical positioning to hold the geoid at a fixed level with respect to the tide gauges (see §19.1); but this makes life difficult for other applications. For the purpose of vertical movement investigations, it is better to accept the fact that the geoid varies with time, and learn to live with it.

### 26.3. Vertical displacement profiles

Of the sources of information discussed in §26.1, the point tilt measurements are of little value, the techniques using observations to extraterrestrial objects are not accurate enough yet, the gravity variations have problems with interpretation, and lake level data are scarce. This leaves the relevelled segments, combined with sea level records wherever possible, as the prime data for vertical movement detection. In many cases, relevelled segments are strung together making a continuous profile. This situation occurs when a whole levelling line is relevelled, and this is commonly recognized as the most favourable setup.

If the first levelling is carried out at epoch  $\tau_1$  and the second at  $\tau_2$ , the *relative vertical displacement* of common bench marks can be simply plotted as the cumulative relative *displacement profile*—see the upper portion of FIG. 6. Since the relevelled line is generally not straight, it is sometimes advisable to plot the profile in a perspective view. For illustration, this has been done in the lower portion of FIG. 6. If the relevelled line is connected to tide gauge  $A$ , for which the record from the period  $\tau_1, \tau_2$  is available, then the relative vertical displacement can be converted to *absolute vertical displacement* by taking into account the absolute displacement  $\delta H_A$  of  $A$ , as indicated by the tide gauge record—see FIG. 7. It should be clear that this

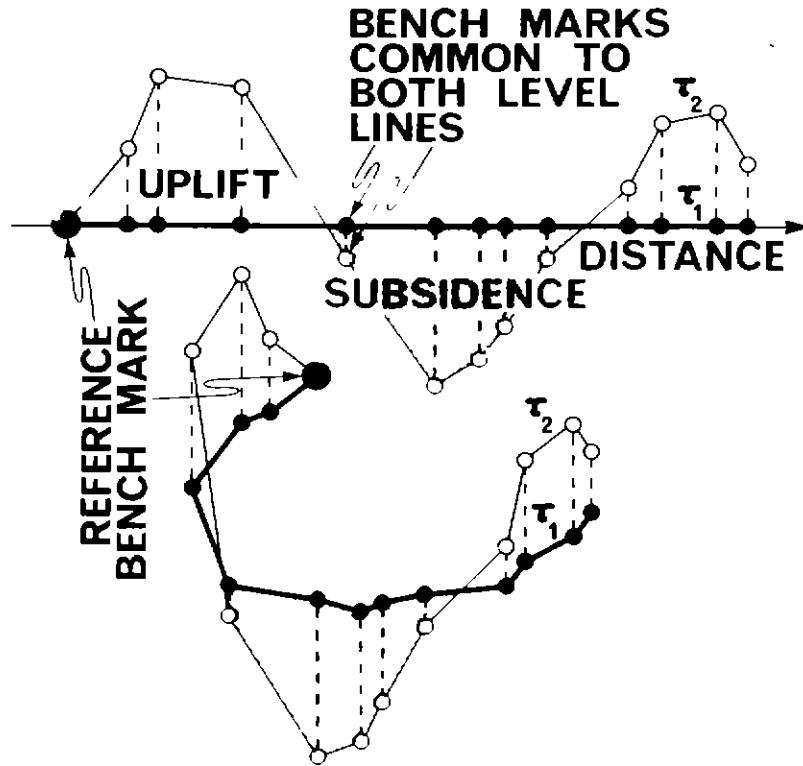


FIG. 26.6. Relative vertical displacement profile.

displacement of the tide gauge  $A$ , assumed to be linear in time, during the period  $\tau_1, \tau_2$ , is given as

$$\delta H_A(\tau_1, \tau_2) = -(c_E - r_E)(\tau_2 - \tau_1), \quad (26.11)$$

where  $c_E$  was defined in §19.1, and  $r_E$  is the eustatic water rise rate; note the negative sign.

As pointed out in §26.1, the relevelled elevation differences are burdened with errors. Not all the variations in the profiles depict the vertical displacement of bench marks; it is, therefore, recommended to plot not only the accumulated displacements  $\delta H_n$  but also the accumulated standard deviations  $\sigma_{\delta H_n}$  of the displacements (cf. FIG. 7). Since  $\delta H_n$  is given by the obvious formula (cf. (1))

$$\delta H_n = \delta H_A + \sum_{i=1}^n \delta \Delta H_i = \delta H_A + \sum_{i=1}^n \Delta H_i(\tau_1) - \sum_{i=1}^n \Delta H_i(\tau_2), \quad (26.12)$$

then its standard deviation  $\sigma_{\delta H_n}$  can be written as

$$\sigma_{\delta H_n}^2 = \sigma_{\delta H_A}^2 + \mathbf{u} \mathbf{C}_{\delta \Delta H} \mathbf{u}^T = \sigma_{\delta H_A}^2 + \mathbf{u} (\mathbf{C}_{\Delta H(\tau_1)} + \mathbf{C}_{\Delta H(\tau_2)}) \mathbf{u}^T, \quad (26.13)$$

where  $\mathbf{u}$  is once more a vector of  $(n - 1)$  ones. When the covariances are expected to be close to zero (statistically independent determinations of  $\Delta H$ ), which is the assumption usually adopted by necessity in practice, then (13) degenerates to

$$\sigma_{\delta H_n}^2 = \sigma_{\delta H_A}^2 + \sum_{i=1}^n [\sigma_{\Delta H_i(\tau_1)}^2 + \sigma_{\Delta H_i(\tau_2)}^2]. \quad (26.14)$$

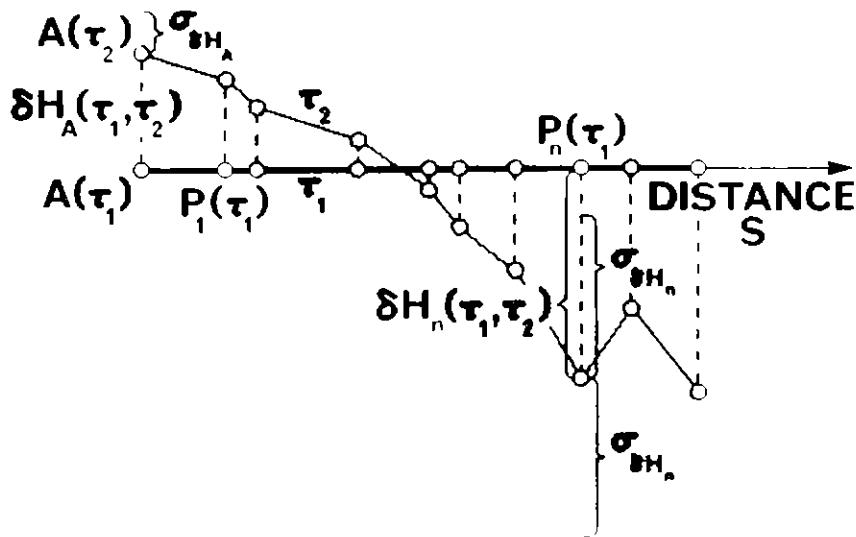


FIG. 26.7. Absolute vertical displacement profile with error bounds.

If, in addition, each levelling was carried out to a specific accuracy of  $\sigma_1^2$  (corresponding to a unit distance—cf. §19.2), then the a posteriori estimate  $\hat{\sigma}_1^2$  of this  $\sigma_1^2$  can be used to further simplify the above equation to

$$\sigma_{\delta H_n}^2 = \sigma_{\delta H_A}^2 + [\hat{\sigma}_1^2(\tau_1) + \hat{\sigma}_1^2(\tau_2)] \sum_{i=1}^n \Delta S_i$$

or

$$\boxed{\sigma_{\delta H_n}^2 = \sigma_{\delta H_A}^2 + S_n [\hat{\sigma}_1^2(\tau_1) + \hat{\sigma}_1^2(\tau_2)]}, \quad (26.15)$$

where  $S_n$  is the accumulated distance of the  $n$ th bench mark  $P_n$  from  $A$ .

It is of interest to realize that if the results of the two levellings of the same segment ( $\Delta H_i(\tau_1)$  and  $\Delta H_i(\tau_2)$ ) are positively statistically dependent, then the error in  $\delta\Delta H_i$  is smaller. Denoting the covariance between the two results by  $\sigma_i(\tau_1, \tau_2)$  and their variances by  $\sigma_i^2(\tau_1)$ ,  $\sigma_i^2(\tau_2)$ , we can write for the correlation coefficient  $\rho_{12}$ ,

$$\rho_{12} = \sigma_i(\tau_1, \tau_2) / (\sigma_i(\tau_1), \sigma_i(\tau_2)). \quad (26.16)$$

On the other hand, the covariance law yields

$$\sigma_{\delta\Delta H_i}^2 = \sigma_i^2(\tau_1) - 2\sigma_i(\tau_1, \tau_2) + \sigma_i^2(\tau_2). \quad (26.17)$$

Assuming now, without any loss of generality,  $\sigma_i(\tau_1) = \sigma_i(\tau_2) = \sigma_i$ , we have

$$\boxed{\sigma_{\delta\Delta H_i}^2 = 2\sigma_i^2(1 - \rho_{12})}, \quad (26.18)$$

which is obviously  $(1 - \rho_{12})$ -times smaller than if there was no statistical dependence. Based on the established statistical dependence between forward and backward runnings (see §19.3), it is reasonable to expect the degree of statistical

dependence between two levellings over the same ground to be quite high. This statistical dependence is caused by the common systematic effects that depend on height, height gradients, type of soil, azimuth of segments, latitude, etc. Thus the vertical displacements obtained from relevelled segments are likely to be significantly more accurate than the levelled heights.

The question remains, however, as to what effect the residual refraction (see §19.2) may have on  $\delta\Delta H_i$ . To investigate this effect, let us rewrite (19.12), using the slope of the terrain  $\beta \doteq \delta l/\Delta S$ , as

$$\delta H_R \doteq A \Delta t \beta \Delta S^3. \quad (26.19)$$

The refraction effect on  $\delta\Delta H_i(\tau_1, \tau_2)$  is then obtained approximately as the difference

$$\delta(\delta\Delta H_i)_R \doteq A \beta_i [\Delta t(\tau_2) \Delta S_i^3(\tau_2) - \Delta t(\tau_1) \Delta S_i^3(\tau_1)]. \quad (26.20)$$

Clearly, if neither the sight lengths  $\Delta S_i$  nor the temperature gradients  $\Delta t$  have changed significantly from one levelling to the other, then the effect is negligible. Fortunately, the latter is usually the case on steep slopes (for large  $\beta_i$ ), where the maximum possible sight length is dictated by the slope rather than other considerations. On the other hand, if a systematic change in the lengths  $\Delta S$  occurs, e.g., dictated by a change in field procedures, then a systematic effect on height-difference differences may accrue.

Obviously, if only two levellings of the same line are available, then the only thing that can be done is to interpret the indicated vertical displacements along the levelling line as resulting from a non-accelerated, or *linear vertical movement* that has taken place during the period between  $\tau_1$  and  $\tau_2$ . For the time being, let us assume the movements to be linear (it will be shown later how the accelerated, or non-linear, movements are treated), and let us proceed toward a slightly more complicated case of loops. If a loop has been levelled in two parts *A* to *B* and *B* to *A* at two different epochs  $\tau_1$  and  $\tau_2$ , then one no longer can expect the height differences around the loop to add up to zero, i.e., one can no longer force the levelling to close. The misclosure, called the *kinematical loop misclosure*, is indicative of the amount of differential movement between *A* and *B* that has taken place in the period  $\tau_1, \tau_2$ . Nothing more, however, can be inferred from it about the movements along the loop. This becomes possible only when the loop, or parts of the loop, have been relevelled. If the loop, as a whole, has been levelled twice at epochs  $\tau_1, \tau_2$ , then the loop can be treated as a closed profile in the same way as the individual lines were treated above. The only difference is that the two end points of such a loop profile, which both correspond to the selected reference point in the loop, show zero relative displacement.

The situation becomes more complicated when individual lines of the loop have been levelled at different epochs. One simple example is displayed in FIG. 8. While the first levelling carried out at epoch  $\tau_1$  produces a misclosure that can be

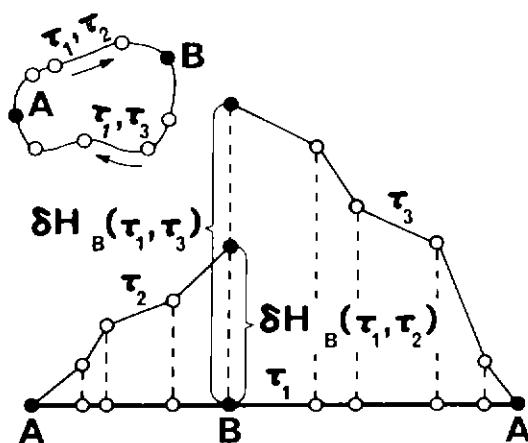


FIG. 26.8. Loop levelled at three different epochs and plot of relative displacements.

distributed around the loop, the relevelings done at  $\tau_2$  and  $\tau_3$  do not, because they cannot be put together. The height difference  $\Delta H_{AB}(\tau_2)$  is generally different from  $\Delta H_{AB}(\tau_3)$  (see FIG. 8). Clearly, under the assumption of linearity of the movements, the following condition should be satisfied,

$$\frac{\delta \Delta H_{AB}(\tau_1, \tau_3)}{\tau_3 - \tau_1} - \frac{\delta \Delta H_{AB}(\tau_1, \tau_2)}{\tau_2 - \tau_1} = v_B - v_A = \delta v_{AB} = \text{const.}, \quad (26.21)$$

where the *constant vertical velocity* of A is denoted by  $v_A$ , and that of B by  $v_B$ . If the first levelling of the loop was adjusted to close, then the difference  $\delta \Delta H_{AB}(\tau_1, \tau_2) - \delta \Delta H_{AB}(\tau_1, \tau_3)$  is nothing less than the misclosure accumulated during the period  $\tau_2, \tau_3$ . Cases of more than two breaks in the loop, and cases of loops consisting of levelling lines of different orders, are left to the reader.

The idea of using kinematical misclosures for the adjustment of relative velocities can be readily extended to cover the entire network of relevelled loops [KORHONEN, 1961]. This topic will be treated in detail in the next section. What must be pointed out here is that the standard, stationary adjustment of vertical positions (cf. §19.2) should not be attempted on loops which have not been observed at one epoch each. Such an adjustment would be justifiable only when the vertical velocity differences involved can all be assumed to be equal to zero. In the opposite case, the adjustment results in a distortion of heights which cannot then be regarded as referring to a definite epoch.

The obvious question then arises: If a loop has been completely levelled at two different epochs, can we simply adjust the heights for the first epoch, do the same for the second epoch, and derive the vertical displacements as differences of the two sets of heights determined in this way? The answer is yes! There is, however, a drawback to this approach: one has to be very careful to formulate the error model properly. The standard formulation invariably leads to an overestimation of errors (i.e., an underestimation of accuracy) of the resulting displacements. The reasons for this behaviour will be shown in the next chapter in the context of horizontal networks.

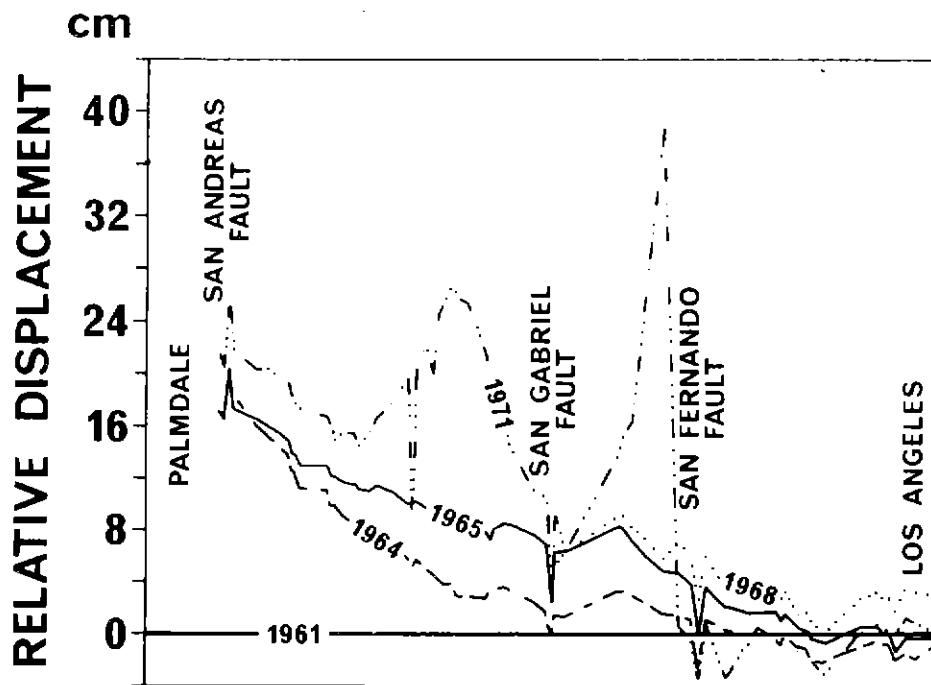


FIG. 26.9. Spatial profiles of preseismic and coseismic vertical displacements.

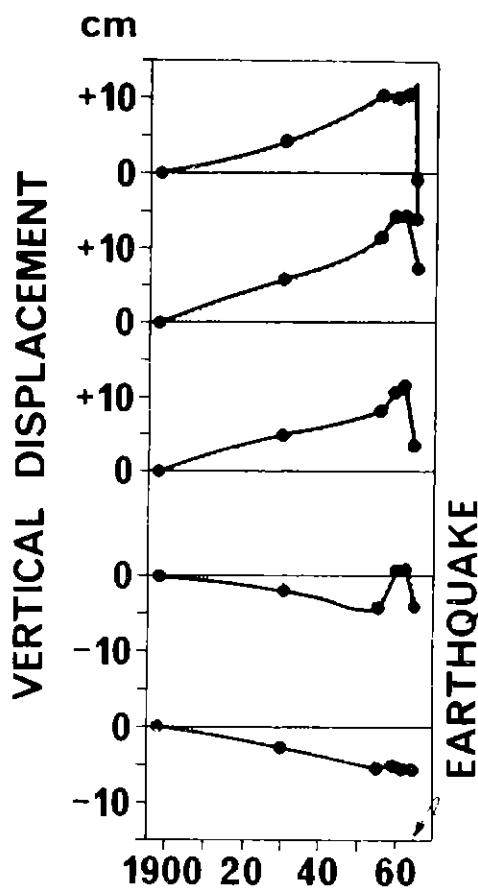


FIG. 26.10. Temporal profiles of preseismic and coseismic vertical displacements.

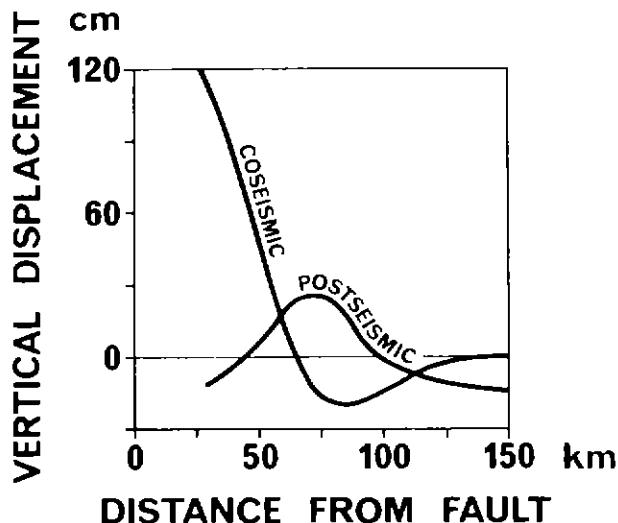


FIG. 26.11. Pattern of coseismic and postseismic vertical displacement.

In regions where the movements are known or suspected to be non-linear, i.e., when we are faced with *accelerated movements*, their determination requires levellings that have been carried out at more than just two epochs. This need arises primarily in earthquake-prone areas. Well developed geodetic programmes for seismically active areas exist in Japan and, to a certain extent, in the U.S. [RIKITAKE, 1976]. These programmes, with the overall aim of helping to understand the earthquake triggering processes, concentrate on measuring preseismic, coseismic, and postseismic movements along tectonic plate boundaries (cf. §8.3).

Results of repeated levellings for the determination of accelerated movements may be plotted in a set of *spatial profiles*. FIG. 9 shows one set of such profiles, observed over the period 1961 to 1971, spanning the earthquake of magnitude 6.4 that occurred in San Fernando, California, on February 9, 1971 [CASTLE ET AL., 1974]. The coseismic and preseismic displacements are clearly visible on these profiles.

Alternatively, *temporal profiles*, instead of spatial, may sometimes be used to advantage. FIG. 10 contains a series of such temporal profiles for five points straddling the location of the 1964 Niigata, Japan, earthquake of magnitude 7.5 [TSUBOKAWA ET AL., 1968]. The profiles show rather nicely the (precursory) vertical displacements preceding the earthquake.

For comparison, let us have a look at the character of coseismic and postseismic vertical displacements of a strike-slip earthquake. FIG. 11 shows the observed movements during and after an earthquake of magnitude 8.2 (at Nankaido, Japan, in 1946), according to NUR AND MAVKO [1974].

## 26.4. Areal modelling of vertical movements

Let us return once more to the linear movements and formulate the *mathematical model for the kinematical adjustment of a height network* mentioned in the previous section. Assuming the observed relative velocities  $\delta v_{ij}^{(0)}$ , which can be written as

(cf. (21))

$$\delta v_{ij}^{(0)} = \frac{\delta \Delta H_{ij}(\tau_1, \tau_2)}{\tau_2 - \tau_1} = \frac{\Delta H_{ij}(\tau_2) - \Delta H_{ij}(\tau_1)}{\tau_2 - \tau_1}, \quad (26.22)$$

to be statistically independent, the inverse of their variances can be used for weights. The variances are given by the formula (see (15))

$$\sigma_{\delta v_{ij}}^2 = \frac{\Delta S_{ij}}{(\tau_2 - \tau_1)^2} [\hat{\sigma}_1^2(\tau_1) + \hat{\sigma}_1^2(\tau_2)] = \Delta S_{ij} \hat{\sigma}_1^2(v). \quad (26.23)$$

We thus have everything needed to write the *observation equation for a relevelled segment*,

$$r_{ij}^{\delta v} = v_j - v_i - \delta v_{ij}^{(0)}, \quad (26.24)$$

which is linear and explicit in the observable, and the adjustment of a completely relevelled network can be carried out. Sea level information, i.e., the vertical velocity  $v_i$  of a tide gauge (cf. (11)), supplies the *observation equations for a point velocity*:

$$r_i^v = v_i - \frac{\delta H_i(\tau_1, \tau_2)}{\tau_2 - \tau_1} = v_i + (c_E - r_E), \quad (26.25)$$

which can be easily merged with the observation equations for levelling data. These velocities are weighted in the usual way by means of their variances:

$$\sigma_{v_i}^2 = \sigma_{\delta H_i}^2 / (\tau_2 - \tau_1)^2 = \sigma_{c_E}^2 + \sigma_{r_E}^2. \quad (26.26)$$

A typical example of such a kinematically adjusted network is the Finnish levelling network [KÄÄRIÄINEN, 1953]. Although the Finnish height network serves primarily as a height control, it has also become a powerful tool for studying the Fennoscandian postglacial rebound (cf. §8.2). The decision to relevel the network completely every forty years was taken earlier this century. Similar decisions have been made by other countries, notably England [EDGE, 1959]. The selection of the frequency of relevellings is not crucial as long as the assumption of linearity of movements can be substantiated.

Let us now focus our attention on the impact of the kinematical adjustment on the displacement profiles along the individual level lines, a problem that parallels the back distribution of residuals in levelling treated in §19.2. Once the adjusted velocities  $\hat{v}$  of the junction points are known, the velocities of the intermittent bench marks can be derived. This is mostly done assuming statistical independence of the relevelled height differences. The back distribution of the kinematical residuals  $r_{ij}^{\delta v}$  is

a simple matter of linear interpolation with distance, according to (19.35). The error estimates of the adjusted velocities of individual bench marks are then obtained through combining the error estimates in the adjusted velocities of junction points  $P_1, P_2$  (see FIG. 19.12) with the propagation of errors along the profile (cf. (19.38)).

It is easy to see that once the velocities of both the junction points as well as points within the levelling lines are adjusted, they can be plotted as a discrete velocity field. If there are physical grounds for believing that the movements represented by these velocities are not only linear in time but that they also vary smoothly with location, then *spatial prediction of vertical velocities* at other points in the area may be attempted. The usual way of doing this is by hand contouring. An example of such predictions, based on a kinematical adjustment of a relevelled network, is shown in FIG. 12 [HOLDAHL AND MORRISON, 1974].

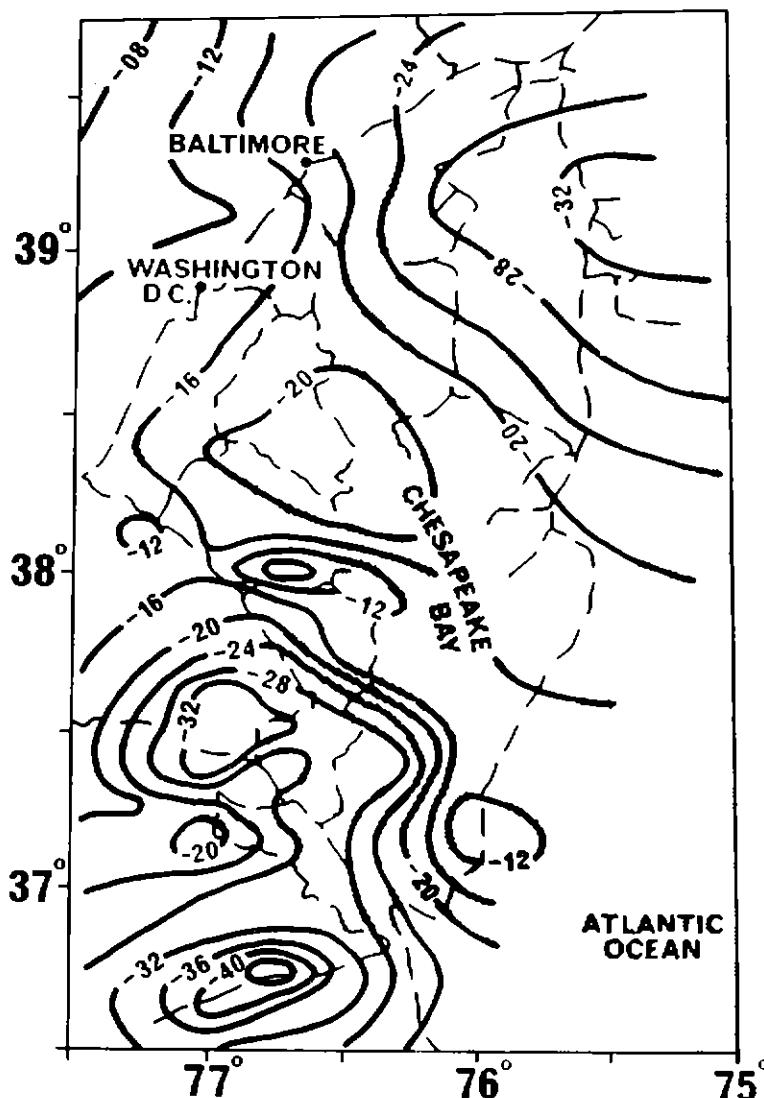


FIG. 26.12. Prediction of spatially smooth linear movements in the Chesapeake Bay, U.S.A., area. Contours in centimetres per century.

Clearly, the prediction through manual contouring can be replaced by an automated process: the contours can be drawn by computer. More fundamentally, though, a *vertical-velocity surface* can be fitted to the discrete velocity field. Here, one should realize that there is no need for an interpolation (see §14.2), i.e., for the requirement that the velocity surface pass exactly through all the discrete points representing the velocities  $v_i$ , since these velocities are determined only to a limited degree of accuracy, characterized by the standard deviations  $\sigma_{v_i}$ . Hence an approximation method should be used, instead of interpolation, to predict the velocity surface. Once more, the simplest procedure is that of surface fitting using least-squares regression. Since this procedure has already been shown in other examples (see §22.3, §24.1, §25.3), we shall not dwell on it here.

It is interesting to realize that if a smooth velocity surface can be regarded as a good enough representation of the actual vertical movements, then it is no longer necessary to require the relevelled segments to be strung together into lines and these lines into a network. Such a smooth velocity surface may be constructed from disconnected, *scattered relevelled segments*. Each segment is then treated as a tilt element, i.e., as a horizontal gradient of velocity between the two end points of the segment. To show how this can be done, let us assume that the velocity surface is wanted in the form

$$v(x, y) = \sum_{i=1}^n \phi_i(x, y) c_i = \tilde{\Phi}^T(x, y) \mathbf{c}, \quad (26.27)$$

where  $\phi_i$  are some selected base functions (cf. §14.2), and  $x, y$  are taken to be in a local coordinate system prescribed, for instance, by (22.65). Then for each relevelled segment  $P_i, P_j$ , the following observation equation [VANÍČEK AND CHRISTODULIDES, 1974] can be written:

$$\delta v_{ij} = v(x_j, y_j) - v(x_i, y_i) = (\tilde{\Phi}^T(x_j, y_j) - \tilde{\Phi}^T(x_i, y_i)) \mathbf{c}, \quad (26.28)$$

or, in a more compact fashion,

$$\boxed{\delta v_{ij} = \widetilde{\Delta \Phi}^T(x_i, y_i, x_j, y_j) \mathbf{c}.} \quad (26.29)$$

It is interesting to note that this formulation is not dissimilar to a finite difference representation of a partial differential equation of first order.

When sufficient relevelled segments are available, distributed so that the selected base functions are linearly independent, the coefficients  $\mathbf{c}$  can be estimated using the least-squares technique. The weight matrix  $P_i$ , again usually considered diagonal, is assembled from the inverse values of variances computed for each relevelled segment from (23). The normal equations then read

$$\widetilde{\Delta \Phi} P_i \widetilde{\Delta \Phi}^T \mathbf{c} = \widetilde{\Delta \Phi} P_i \delta \mathbf{v}. \quad (26.30)$$

The question now arises: How can the relevelled segments be combined with other kinds of information for the construction of the velocity surface? Obviously, the lake

level tilts can be used in precisely the same manner as the relevelled segments, i.e., as tilt elements: they yield observation equations of the same type (29). The only difference is in weighting. The variance of the lake tilt element is given simply as

$$\sigma_{\delta v_{ij}}^2 = \sigma_{v_i}^2 + \sigma_{v_j}^2, \quad (26.31)$$

where  $\sigma_{v_i}$ ,  $\sigma_{v_j}$  are standard deviations of the linear trends determined from the two gauge records.

The situation is different with sea level records. Every sea tide gauge supplies one observation equation of the kind shown earlier (27), where the observed velocity is given by (25). The variance of the observation is spelled out by (26). The two systems of observation equations, i.e., those for tilt elements and those for point velocity, are then combined, following the procedure described in §14.5, to arrive at

$$[\widetilde{\Delta\Phi}^T | \tilde{\Phi}^T]^T c = [\delta v | v]^T. \quad (26.32)$$

The appropriate weight matrix  $P$  for the corresponding system of normal equations is evidently

$$P = \begin{bmatrix} P_t & \mathbf{0} \\ \mathbf{0} & -P_v \end{bmatrix}, \quad (26.33)$$

where  $P_v$  is the weight matrix for the point velocities. The augmented normal equations then read

$$[\widetilde{\Delta\Phi} | \tilde{\Phi}] P [\widetilde{\Delta\Phi}^T | \tilde{\Phi}^T]^T c = [\widetilde{\Delta\Phi} | \tilde{\Phi}] P [\delta v | v]^T. \quad (26.34)$$

As we know from §12.3, the covariance matrix  $C_c$  of the resulting coefficients  $c$  is obtained simply as the appropriately scaled inverse of the matrix of normal equations. It can then be used in the evaluation of the *standard deviation of the velocity surface* at any point  $(x, y)$ . From (27), the covariance law yields:

$$\sigma_v(x, y) = \sqrt{\tilde{\Phi}^T(x, y) C_c \tilde{\Phi}(x, y)} \quad (26.35)$$

For comparison, FIG. 13 shows the vertical velocity surface and its standard deviation computed by means of this method, using approximately the same data as those used in FIG. 12. Two-dimensional, algebraic functions up to the fourth degree were used for the base [VANÍČEK AND CHRISTODULIDES, 1974]. Other base functions may of course be used; HOLDAHL AND HARDY [1979], for instance, used quadrics.

When the movements cannot be assumed to be linear, for example in tectonically active areas, the above approach only yields some average velocities. If, in addition, the survey epochs are too scattered, then the results are practically useless. The technique can be generalized, however, to also cope with accelerated movements. Such a generalization is achieved by expanding the model to include time. It is then

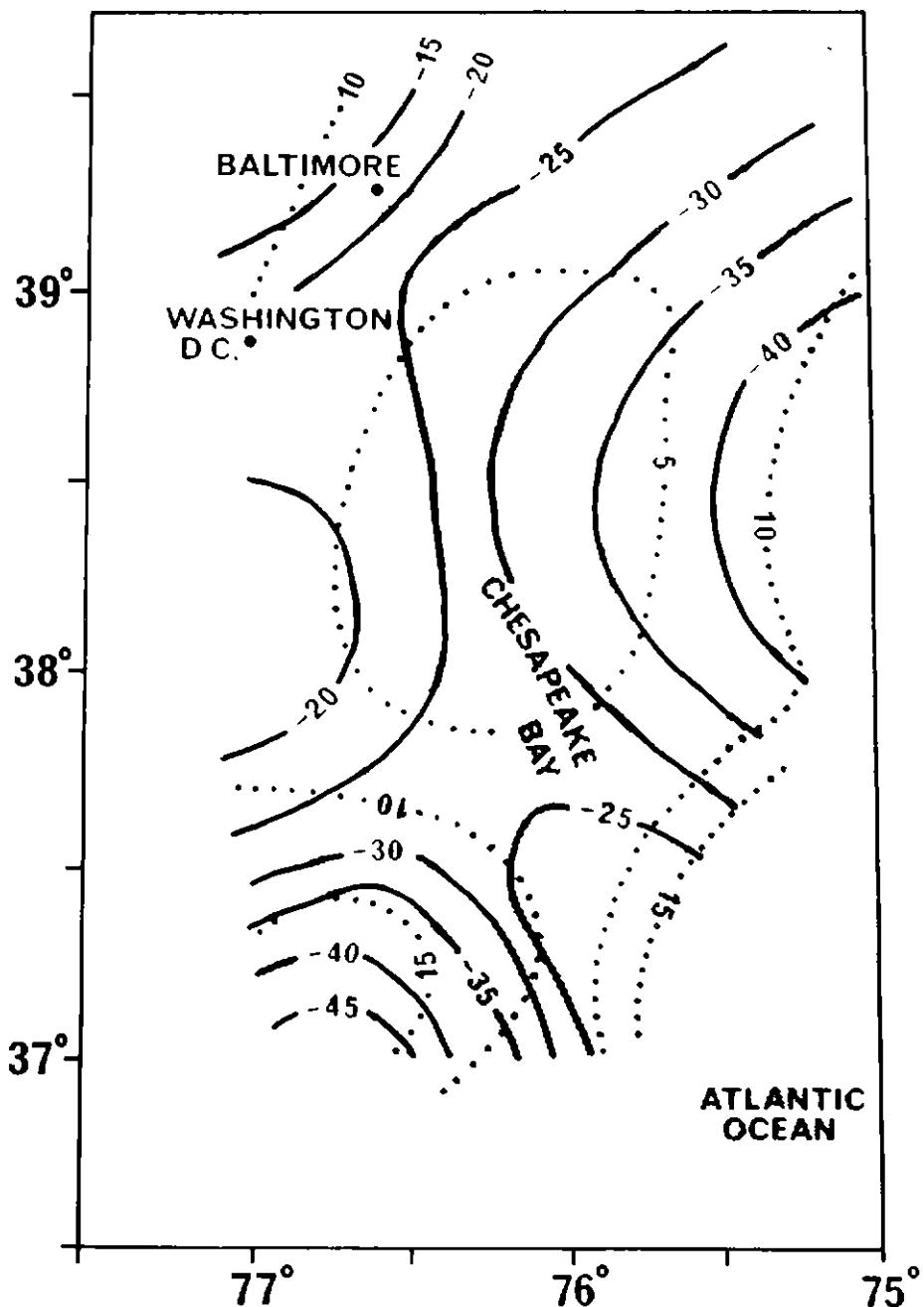


FIG. 26.13. Least-squares prediction of linear movements in the Chesapeake Bay, U.S.A., area. Standard deviation contours dotted. Contours in centimetres per century.

expedient to seek vertical displacements rather than velocities. Denoting the vertical displacement of point  $(x, y)$  during the period  $(\tau_0, \tau)$  by  $u(x, y, \tau)$ , and assuming

$$u(x, y, \tau_0) = 0 \quad (26.36)$$

everywhere, one can write the *four-dimensional model for vertical displacement* as, for instance,

$$u(x, y, \tau) = \mathbf{u}^T \tilde{\Phi}(x, y) \mathbf{T}^T(\tau) \tilde{\mathbf{c}} \mathbf{u}. \quad (26.37)$$

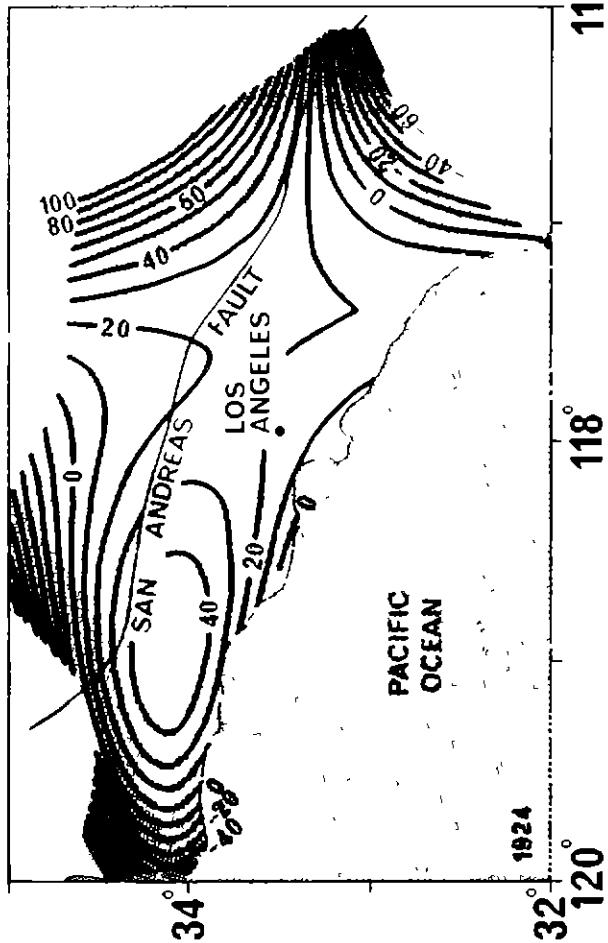
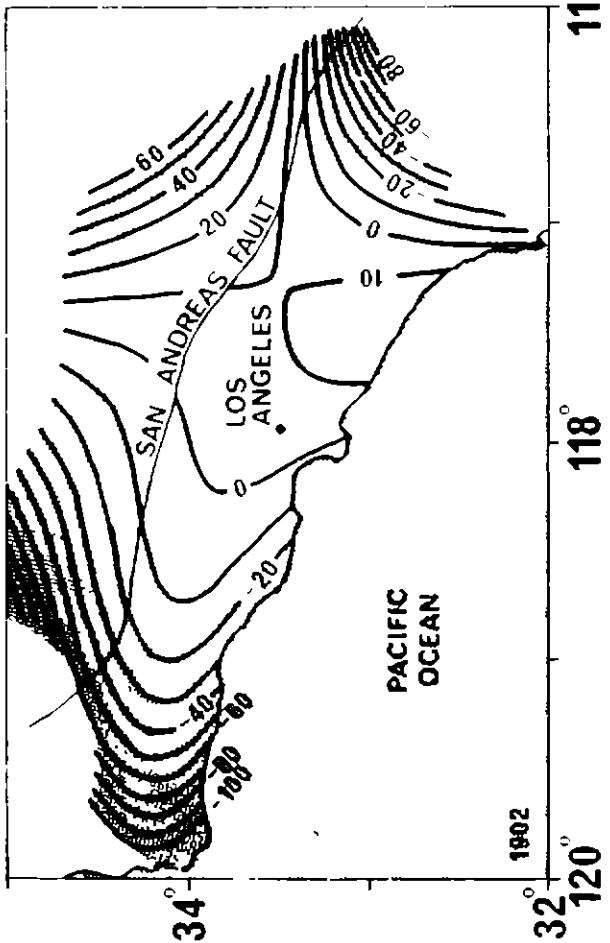
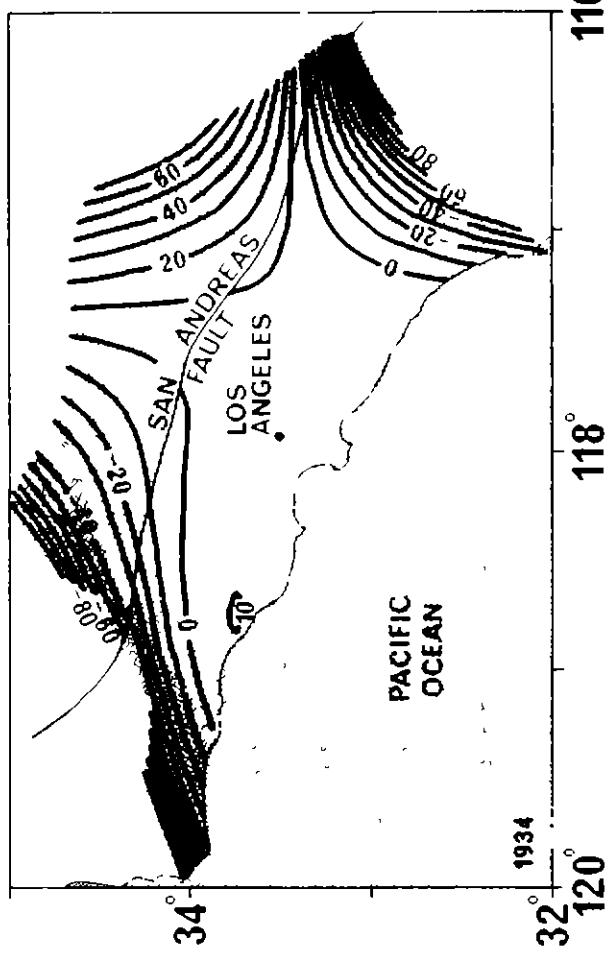
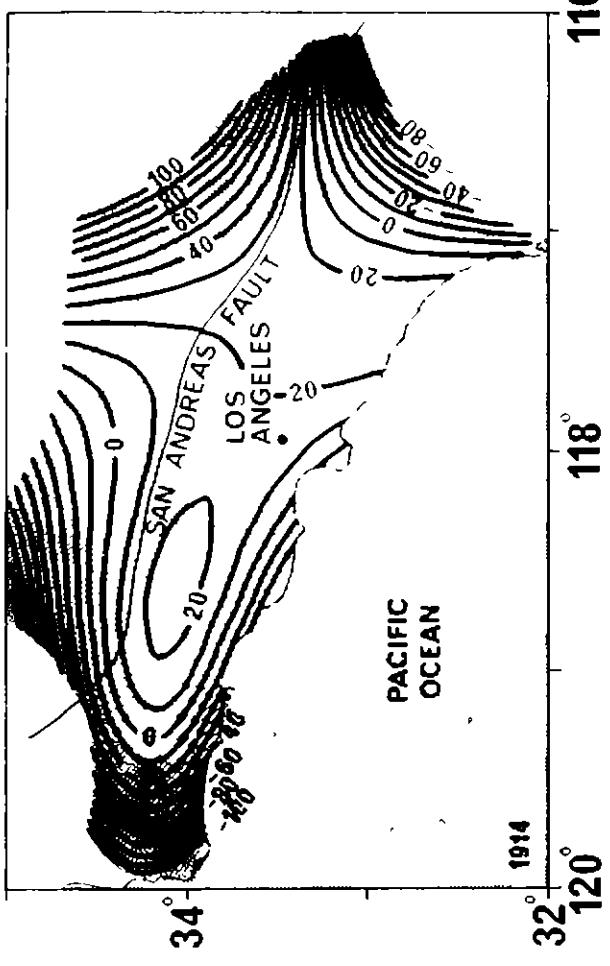


FIG. 26.14. Least-squares prediction of accelerated movements for Southern California. Standard deviation surfaces are shaded. Contours in centimetres per century.

Here  $\mathbf{T}$  is an  $m_T$  vector of base functions of time such that

$$\mathbf{T}(\tau_0) = \mathbf{0}, \quad (26.38)$$

$\tilde{\Phi}$  is an  $m_\Phi$  vector of base functions in space,  $\tilde{c}$  is an  $m_T \times m_\Phi$  matrix of coefficients to be determined, and  $\mathbf{u}$  is again an  $m_\Phi$  vector of ones. From levelling we get only the difference of the vertical displacements of the two end points  $P_i, P_j$  of the relevelled segment during the period  $(\tau_1, \tau_2)$ , i.e.,

$$\begin{aligned} \delta u_{ij}(\tau_1, \tau_2) &= u(x_j, y_j, \tau_2) - u(x_j, y_j, \tau_1) - u(x_i, y_i, \tau_2) + u(x_i, y_i, \tau_1) \\ &= \Delta H_{ij}(\tau_2) - \Delta H_{ij}(\tau_1) = \delta \Delta H_{ij}(\tau_1, \tau_2). \end{aligned} \quad (26.39)$$

Thus as the reader can derive, the mathematical model for one relevelled segment becomes

$$\delta u_{ij}(\tau_1, \tau_2) = \mathbf{u}^T [\tilde{\Phi}(x_j, y_j) - \tilde{\Phi}(x_i, y_i)] [\mathbf{T}^T(\tau_2) - \mathbf{T}^T(\tau_1)] \tilde{c} \mathbf{u}. \quad (26.40)$$

Written in a more compact way,

$$\begin{aligned} \delta u_{ij}(\tau_1, \tau_2) &= \mathbf{u}^T \tilde{\Delta \Phi}(x_i, y_i, x_j, y_j) \Delta \mathbf{T}^T(\tau_1, \tau_2) \tilde{c} \mathbf{u} \\ &= \mathbf{B}^T(x_i, y_i, x_j, y_j, \tau_1, \tau_2) \mathbf{c}, \end{aligned} \quad (26.41)$$

where  $\mathbf{c}$  is an  $(m_T \times m_\Phi) \times 1$  vector of stacked up columns of the  $\tilde{c}$  matrix, and  $\mathbf{B}$  is an  $(m_T \times m_\Phi) \times 1$  vector representing one column of the Vandermonde matrix of the product of the two base function systems. For such four-dimensional modelling, it is advantageous to have the relevelled segments evenly scattered in both space and time.

Merging the above observation equations with observation equations describing the *accelerated variations of sea level*, i.e.,

$$u_A(\tau_1, \tau_2) = \mathbf{u}^T \tilde{\Phi}(x_A, y_A) \Delta \mathbf{T}^T(\tau_1, \tau_2) \tilde{c} \mathbf{u} = \mathbf{D}^T(x_A, y_A, \tau_1, \tau_2) \mathbf{c}, \quad (26.42)$$

we obtain the complete model (cf. §14.5). The model is then solved for  $\mathbf{c}$ , which, in turn, are used to predict the uplift for any point  $(\phi, \lambda, \tau)$  in space and time. The standard deviation of an uplift predicted in this way is evaluated along the same lines as in the case of linear movements. FIGS. 14 show areal displacements, predicted for southern California, using the above method [VANÍČEK ET AL., 1979]. The base used in this particular model is algebraic functions up to degree two in  $y$ , three in  $x$ , and three in  $\tau$ . The reference epoch is  $\tau_0 = 1897$ .

## CHAPTER 27

# DETECTION OF HORIZONTAL MOVEMENTS

In general, there is little that distinguishes horizontal from vertical movements except for their directions and their magnitudes: horizontal movements can be much larger. Thus most of the statements about vertical movements made in the introduction to Chapter 26 will be equally valid here. Naturally, the layout of the sections in this chapter also mirrors, to a large extent, that of the previous chapter. The only significant difference is in the treatment of the reference datums for the two kinds of movements. While the gravity field supplies the height datum, the horizontal datum does not have to depend on the gravity field at all; understandably, this difference is reflected in the way the movements are treated and, as a result, there is no section equivalent to §26.2 here. The problem of interdependence of position and gravity field variations is replaced by the problem of indeterminacy of various parameters describing the horizontal movements. The question of indeterminacy constitutes the real thread common to all the techniques discussed in this chapter.

The first section provides a discussion and assessment of the available sources of data on horizontal movements. From the geodetic point of view, the most intuitively obvious approach to our subject, based on comparison of horizontal positions determined at different epochs, is analysed in the second section. The third section treats the more direct, yet somewhat more complicated, task of evaluating horizontal displacements without the necessity of adjusting positions first. The last group of models, which seeks local, or differential, characteristics of the movements, is shown in the fourth section. These models, although less familiar to geodesists, are probably the most meaningful ones to geophysicists, whose goal it is to discover the physical causes of the observed movements.

### 27.1. Sources of information on horizontal movements

Much like the vertical data, the ideal horizontal data should be of an areal nature and continuous in time. Unfortunately, there is no technique that provides data of this kind; the existing sources are all a far cry from this ideal. Thus, although the horizontal movements are generally much larger than the vertical, their detection is a more difficult proposition than the detection of vertical movements.

There exist four different sources of data on horizontal movements: (a) repeated point positioning, (b) resurveyed horizontal geodetic networks, (c) monitor configuration, and (d) strain gauging.

(a) *Repeated point positioning*, absolute or relative, is the most obvious technique to use. Unfortunately, however, as we have seen in Chapters 15 and 16, the accuracy of the point positioning techniques for regional and, even more so, for local movements is at present not high enough. This is bound to change in the near future, and plans are afoot to start using point positioning based on observations to extraterrestrial objects for horizontal movement detection [NASA, 1979]. One may argue that for global tectonic plate movements, or for other continual movements of similar kinds, the lack of accuracy may be replaced by longer intervals between determinations. But even for global tectonic movements, which are of the order of several centimetres per year, the existing techniques do not yield conclusive results when used on annual or shorter interval bases, and longer intervals are needed before the accumulated relative displacements reach an observable level. Of the forthcoming techniques, the NAVSTAR (cf. §15.3) is the most promising. Also, laser ranging to high altitude satellites will probably be accurate enough, even using mobile lasers, to become useful in this context [BENDER ET AL., 1979].

The only existing technique with an inherently high enough accuracy is radio-interferometry, discussed in §16.1. Shown in FIG. 1 is the situation when two radio telescopes are used simultaneously. When more radio telescopes are deployed, relative horizontal displacements of these with respect to one arbitrarily selected reference telescope site can be obtained. This operational mode is somewhat similar to the situation encountered in the trilateration of geodetic networks (cf. §18.3).

(b) The *resurveyed horizontal geodetic networks* are the largest available source of data today. This is because the data, i.e., reobserved distances, angles, and azimuths, have been and are being collected for other purposes, chiefly for horizontal position

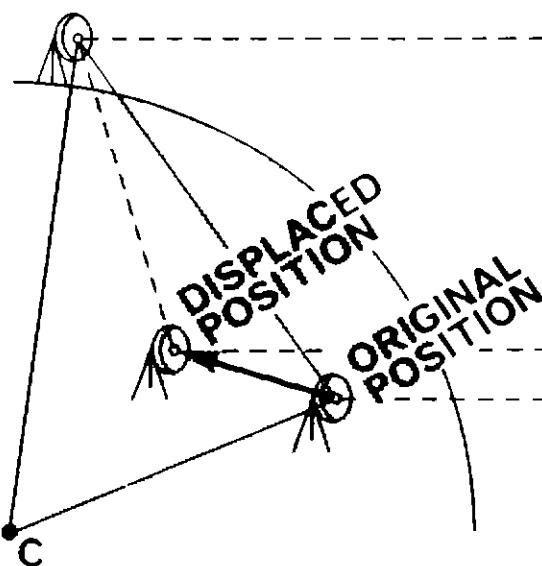


FIG. 27.1. Detection of horizontal movement through radio-interferometry.

control, and are thus readily available for periods of time stretching back for many decades. The price one pays for these freely available data is that they may not be able to give the answer to questions one may want to ask: the vast majority of networks were not, of course, designed for movement detection. Naturally, it is important to have a systematic approach to periodic resurveys of networks. It has been, and very often still is, the custom to resurvey only regions struck by a significant earthquake, where the coseismic displacements are sufficiently large to appreciably change the horizontal coordinates of some control points.

It is in the nature of geodetic networks that the information they yield is of a regional character. FIG. 2 clearly illustrates this fact, showing the horizontal displacements of control points of first, second, and third order in the Tango area of Japan during the period 1900 to 1930, which contains the Tango earthquake of magnitude 7.4 in 1927, according to TSUBOI [1932]. Most examples of detected movements pertain to such larger seismic movements (cf. FIG. 8.20), but the usual relative accuracy of the first-order positions (cf. §7.1) is high enough to detect other kinds of movements as well. For instance, the deformations accumulating along active plate margins are responsible for relative extensions and contractions (strains) up to  $3 \times 10^{-7}$  per year [SAVAGE, 1978]; it is customary to quantify these deformations in terms of dimensionless units, and thus we speak of a maximum order of magnitude of 0.3  $\mu$ strain per year. Such deformations are clearly detectable and, for

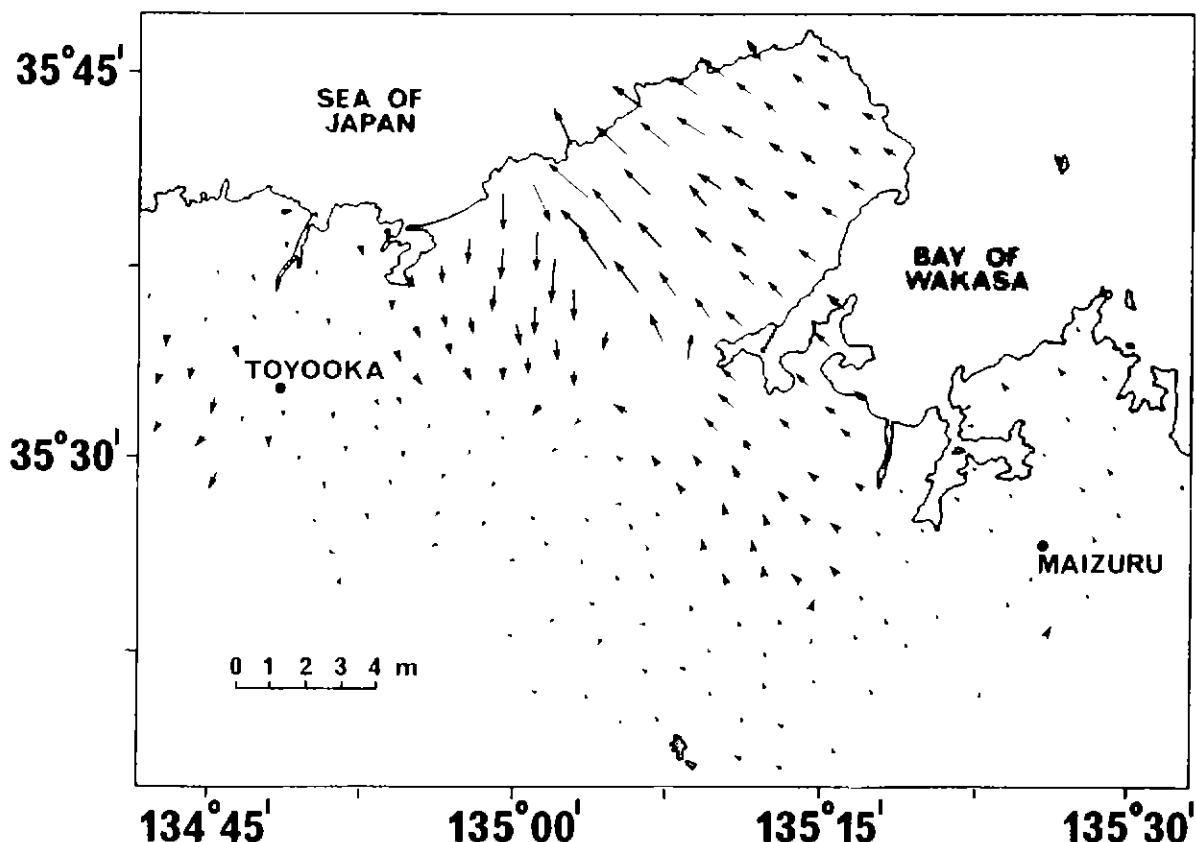


FIG. 27.2. Horizontal displacements of control points in the Tango area, Japan, during the period 1900 to 1930.

an illustration, the reader is referred to FIG. 9 showing horizontal deformations in southern California, U.S.A. It should be noted that if a network is resurveyed several times, then detection of accelerated movements may be attempted; this situation parallels that of the vertical movements.

(c) In regions of special interest, where movements are either suspected or known to be taking place, local *monitoring networks* and other simpler configurations are used to monitor these movements. For obvious reasons, the zones of strain accumulation along plate margins are again the chief but not the sole examples of such regions. As these networks are normally smaller in extent, they can be completely resurveyed more often than the regular geodetic networks. Also, more accurate results are usually obtained simply because more care can be taken of the measurements. CHRZANOWSKI [1980] reports an accuracy of close to  $10^{-6}$  in his monitor networks in Peru, achieved with standard surveying equipment. SAVAGE AND PRESCOTT [1973] claim an accuracy of perhaps one order of magnitude better for the U.S. Geological Survey's monitoring network across the San Andreas fault in California. This accuracy is achieved through distance measurements with a laser instrument, coupled with a very careful sampling of air density along the line of sight. The knowledge of the actual air density, in turn, assists in a much more realistic evaluation of the refraction corrections from (15.39).

The additional advantage of monitoring networks, compared with the regular geodetic ones, is that they can be designed, from the geometrical point of view, to do exactly what is required of them. FIG. 3, for example, shows a chain of monitoring networks designed to detect the areal pattern of strain accumulation along the San Andreas fault system [MEADE AND SMALL, 1966].

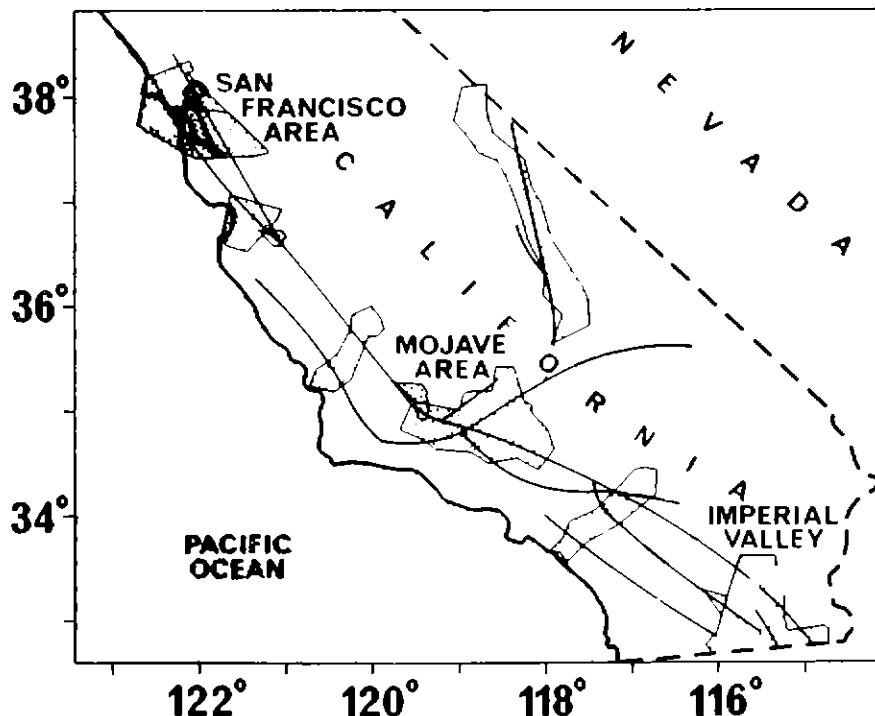


FIG. 27.3. Monitoring networks straddling the San Andreas fault system.

At times, the areal pattern is not the main object of interest. If the interest is in learning the exact rate of lateral slip along an active fault, then a simpler shape for the monitoring figure can be designed. As an illustration, FIG. 4 shows one such configuration used on the San Andreas fault [MEADE, 1973] in which azimuths, angles, and distances were repeatedly measured between 1957 and 1963. An even simpler figure was used by VACQUIER AND WHITEMAN [1973] for essentially the same purpose. Their configuration consisted of three horizontally aligned points across California Bay, and the movement of the points out of alignment was photographically recorded. The technique thus amounted to the registration of changes in the horizontal angle between the points.

(d) When only the rate of extension or contraction, i.e., strain, in one direction is the object of interest, then repeated measurements of the length of an appropriately oriented single line provide the needed information. The radio-interferometrically determined distances mentioned earlier may, in a way, serve as an example of this approach. The main problem with this approach is that the strain determined in this way represents a mean value along the repeatedly observed distance, and thus shorter distances are preferred. A very high accuracy in strain detection is achieved in repeated calibrations of *geodetic base lines*, where the Vaisälä technique, for instance, is reputed to give a relative accuracy in distance determination of better than  $10^{-7}$ , i.e.,  $0.1 \mu\text{strain}$  [BOMFORD, 1971].

When speaking of detection of strain in one direction, one must also mention *strain gauges*. These are instruments specially designed to register temporal changes in a short distance. They are based on various different physical principles the descriptions of which are beyond the scope of this book. Strain gauges vary in length from a few metres to several hundreds of metres. They are the only specialized instruments used in horizontal movement detection, and the only ones that yield continuous records in time. The accuracy of strain gauges is now reaching almost the level of 1 nstrain [LEVINE, 1978], clearly high enough to routinely detect even tidal strain (cf. §25.2). Strain gauges are being used to monitor both the strain accumulation and creep in the regions of interplate tectonic movements. The U.S. Geological

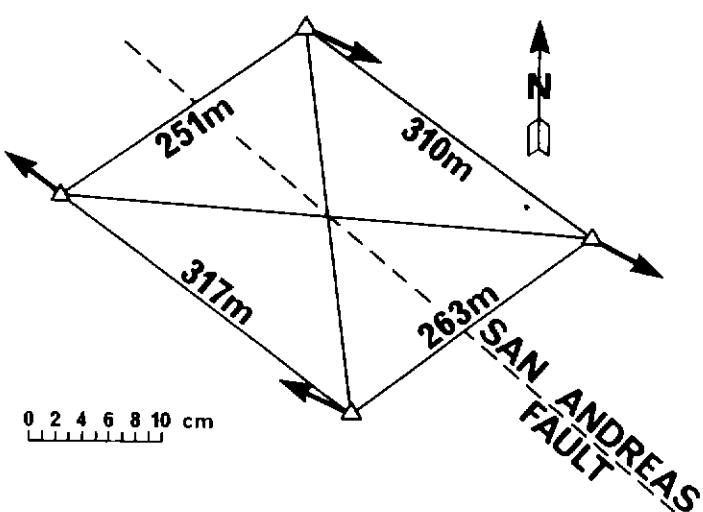


FIG. 27.4. Monitoring figure for detection of slip rate along San Andreas fault near Hollister, California.

Survey work [BURFORD ET AL., 1978] may be cited here as a representative example of scores of strain measurements carried out in different parts of the world.

## 27.2. Comparison of horizontal positions

Turning now to possible analytical techniques for the evaluation of horizontal movements, it is natural to begin with the technique that most geodesists would consider to be the common sense method—*comparison of horizontal positions*. This approach is based on the seemingly straightforward idea that if the horizontal position of a point is determined twice, each time from observations carried out at a specific epoch, then the difference between positions determined in this way reflects the displacement the point has undergone between the two epochs. This notion appears to be straightforward only because it is not complete: what must be added is that it is based on the assumption that the two successive positions are referred to the same coordinate system. A somewhat similar situation was met in §26.2, where the reference system for heights, which is based on the earth's gravity field, had to be shown stable enough to allow the interpretation of observed elevation changes as vertical displacements.

When discussing the concepts of horizontal position comparison, it will be assumed that we are dealing with a whole network of points. Other configurations, such as pairs of points, triangles, etc., may be regarded as the simplest cases of a network. Further, it will be postulated that each observation campaign was carried out almost instantaneously at epochs  $\tau_1$  and  $\tau_2$ . Accordingly, the first batch of observations will be denoted by  $I(\tau_1)$ , or simply by  $I_1$ , and the second by  $I(\tau_2)$ , or  $I_2$ . If the condition of near-simultaneity of observations is not satisfied, then the described approach cannot be used; suitable alternative techniques will be discussed in §27.3. Finally, it will be assumed, for simplicity, that the horizontal positions derived from the observations are expressed in a conformal mapping coordinate system  $(x, y)^M$ , i.e., that the observations needed for the position determination have been properly reduced to a conformal mapping plane.

Under these conditions, the linearized mathematical model developed in §18.2 applies; denoting the corrections to approximate coordinate values  $x^{(0)}$  by  $\delta$ , and the design matrix (cf. (18.13), (18.14), and the observation equation for horizontal angles) by  $A$ , one obtains (cf. 12.7))

$$\mathbf{r}_j = \mathbf{A}_j \boldsymbol{\delta}^{(j)} + \mathbf{w}_j, \quad j = 1, 2. \quad (27.1)$$

Considering then the covariance matrices  $C_i^{(1)}, C_i^{(2)}$ , the least-squares solutions  $\hat{x}^{(1)} = x^{(0)} + \hat{\delta}^{(1)}$  and  $\hat{x}^{(2)} = x^{(0)} + \hat{\delta}^{(2)}$  can be obtained following procedures outlined in §12.2. Also, the covariance matrices  $C_x^{(1)}$  and  $C_x^{(2)}$  are easily derived using the methodology shown in §12.3. This process is nothing less than the routine adjustment of positions as carried out in daily geodetic practice; it is the clear advantage of this approach that standard positioning adjustment techniques are used. An additional advantage accrues if, from the point of view of movement detection, position adjustments improve the observations by forcing the adjusted observations

to conform to the geometry of the network. This may not always be the case, because the requirements on the network for positioning and for movement detection are generally different. For example, a line affected by lateral refraction is to be avoided in a position control network, whereas it makes no difference when displacements are considered, as long as the refraction effect remains more or less stationary.

The *horizontal displacements* are calculated as the differences between the two sets of positions. We can thus write for the vector of horizontal displacements  $\Delta\hat{x}$

$$\Delta\hat{x} = \hat{x}^{(2)} - \hat{x}^{(1)}. \quad (27.2)$$

Since the displacement vector of a point  $P_j$  is given as a pair of numbers

$$v_{x_j} = x_j^{(2)} - x_j^{(1)}, \quad v_{y_j} = y_j^{(2)} - y_j^{(1)}, \quad (27.3)$$

it is expedient to regard the mapping plane as being complex. Then each point  $P_j$  is described by the complex number  $z_j^* = x_j + i y_j$ , and (2) can be rewritten as

$$\Delta\hat{z}^* = \hat{z}^{*(2)} - \hat{z}^{*(1)}. \quad (27.4)$$

Evidently, these vectors of complex numbers have as many components as there are points in the network, and only half as many as have the real vectors in (2).

Now we are ready to look at the problem of indeterminacy of displacements:

(a) To begin with, let us consider a network where both observation sets  $I_1$  and  $I_2$  contain observations of distances, angles or directions, and azimuths. All these observations reflect only the relative positions of adjacent points as well as the orientation of some of the lines at the time of observation. Clearly, the derived positions are only of a relative nature regardless of which coordinate system is used. In other words, the adjusted coordinates can be considered as being referred to one of the points of the network. Let us denote this point as  $P_1$  and write, without any loss of generality,

$$\hat{z}_1^{*(1)} = \hat{z}_1^{*(2)} = 0^*. \quad (27.5)$$

But how do we know that point  $P_1$  has not moved during the period  $(\tau_1, \tau_2)$ ? We do not! There is no way of determining, from the observations  $I_1, I_2$ , whether  $P_1$  has moved with respect to the coordinate system being used. This property, which may be called *indeterminacy in translation*, is a consequence of the invariance of the shape of a horizontal network with respect to a horizontal translation (cf. §18.1—the freedom to select  $\phi_0, \lambda_0$  arbitrarily).

When a particular solution  $\Delta\hat{z}_p^*$  of a position adjustment, where an arbitrary point was assumed fixed, is given, the complete solution  $\Delta\hat{z}_c^*$  is obtained as

$$\Delta\hat{z}_c^* = \Delta\hat{z}_p^* + \mathbf{u}(\Delta x_0 + i\Delta y_0), \quad i = \sqrt{-1}, \quad (27.6)$$

where  $\mathbf{u}$  is a vector of ones of a dimension equal to the number of network points, and  $\Delta x_0, \Delta y_0$  are arbitrary numbers. Regarding the product  $\mathbf{u}(\Delta x_0 + i\Delta y_0)$  as a

complex vector function  $z_0^*(\Delta x_0, \Delta y_0)$ , one gets simply

$$\Delta \hat{z}_c^* = \Delta \hat{z}_p^* + z_0^*(\Delta x_0, \Delta y_0). \quad (27.7)$$

(b) The situation becomes more complicated when azimuths are missing from the observation sets  $I_1$  or  $I_2$ . Let us assume, to begin with, that  $I_1$  does not contain any observed azimuth. Then the solution for position  $\hat{z}^{*(1)}$  is, in addition to being indeterminate with respect to translation, *indeterminate with respect to a rotation* around the origin of the coordinate system (cf. §18.1—the freedom to select  $\alpha_0$  arbitrarily). Thus the complete solution for  $\hat{z}^{*(1)}$  is, in this case,

$$\hat{z}_c^{*(1)} = \hat{z}_p^{*(1)} \exp(-i\psi^{(1)}) + z_0^*(x_0^{(1)}, y_0^{(1)}), \quad (27.8)$$

with three arbitrary constants  $\psi^{(1)}, x_0^{(1)}, y_0^{(1)}$ . Similar equations can be written for the complete solution  $\hat{z}_c^{*(2)}$ , if  $I_2$  does not contain any observed azimuth.

Turning now to horizontal displacements, one obtains

$$\begin{aligned} \Delta \hat{z}_c^* &= \hat{z}_c^{*(2)} - \hat{z}_c^{*(1)} = \hat{z}_p^{*(2)} \exp(-i\psi^{(2)}) - \hat{z}_p^{*(1)} \exp(-i\psi^{(1)}) \\ &\quad + z_0^*(x_0^{(2)}, y_0^{(2)}) - z_0^*(x_0^{(1)}, y_0^{(1)}). \end{aligned} \quad (27.9)$$

Denoting  $\psi^{(2)} - \psi^{(1)}$  by  $\Delta\psi$  (where  $\psi^{(1)}$  or  $\psi^{(2)}$  may be equal to 0, if  $I_1$  or  $I_2$  contains an observed azimuth),  $x_0^{(2)} - x_0^{(1)}$  by  $\Delta x_0$ , and  $y_0^{(2)} - y_0^{(1)}$  by  $\Delta y_0$  yields

$$\Delta \hat{z}_c^* = \Delta \hat{z}_p^* + \hat{z}_p^{*(1)} [\exp(-i\Delta\psi) - 1] + z_0^*(\Delta x_0, \Delta y_0). \quad (27.10)$$

It is illustrative to have a look at possible manifestations of these indeterminacies: FIG. 5 shows two particular solutions, both compatible with the same observations and reducible to zero displacements.

(c) The most complicated case occurs when, in addition to the absence of azimuth, one of either of the observation sets does not contain any observed distance. It is left to the reader to show that the complete solution is then given as

$$\Delta \hat{z}_c^* = [\Delta \hat{z}_p^* + \hat{z}_p^{*(1)} (\exp(-i\Delta\psi) - 1) + z_0^*(\Delta x_0, \Delta y_0)] (1 - \Delta\kappa), \quad (27.11)$$

where  $\Delta\kappa$  is the additional arbitrary parameter which reflects the *indeterminacy in scale*.

Is there anything that can be done to remove these indeterminacies? Nothing, unless additional information about the observed phenomenon is available. For instance, the knowledge that one point, say  $P_j$ , has not moved with respect to a reference frame between the two epochs  $\tau_1, \tau_2$ , i.e.,  $(\Delta \hat{z}_j^*)_c = 0^*$ , leads to the determination of  $\Delta x_0, \Delta y_0$ . This removes the indeterminacy in translation. If one point is held fixed while the displacement of another point is constrained to take place in a prescribed direction, then not only the translation but also the rotation  $\Delta\psi$  can be solved for. When two points are held fixed, all four unknowns can be eliminated; the proof is left to the reader.

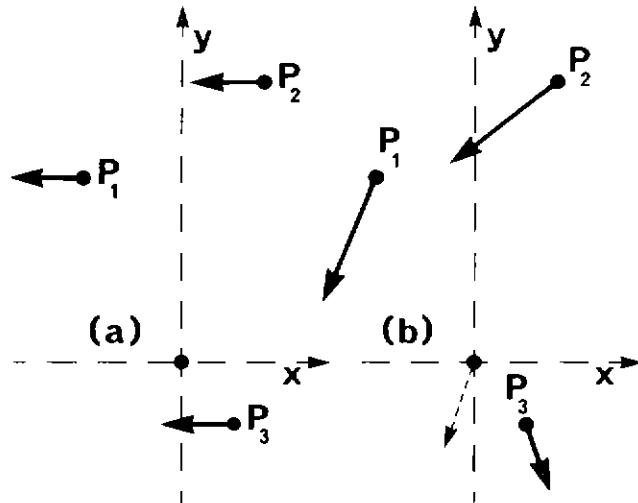


FIG. 27.5. Two particular solutions of the same zero displacement problem. Solution (a) obtained through translation parallel to  $x$ -axis; solution (b) obtained through anticlockwise rotation by  $\pi/7$  around  $(x=0, y=0)$ , and translation shown by the dashed arrow.

Another possibility is to constrain a particular solution by requiring, for example, the sum of all the squares of displacements to be the minimum, i.e.,

$$\min_{\Delta x_0, \Delta y_0, \Delta \psi} \sum_i |\Delta \hat{z}_i^*|_p^2 \Rightarrow \Delta x_0, \Delta y_0, \Delta \psi. \quad (27.12)$$

This is equivalent to

$$\min_{\Delta x_0, \Delta y_0, \Delta \psi} \sum_i (\hat{v}_{x,i}^2 + \hat{v}_{y,i}^2)_p \Rightarrow \Delta x_0, \Delta y_0, \Delta \psi. \quad (27.13)$$

(Note that  $\Delta \kappa$  cannot be determined from this minimization: clearly, only the trivial solution  $\Delta \kappa = 1$  can be obtained.) From a geophysical point of view, however, the imposition of a condition such as (12) does not make much sense, and thus does not solve the indeterminacy problem universally.

Equation (11) provides a handy tool to study the effects of some systematic errors in the observations. When either or both the distances and azimuths are observed both times, then the known average systematic error in their determination can be substituted for  $\Delta \kappa$  and  $\Delta \psi$ , and the effect on computed displacements evaluated. More about systematic errors and their effects may be found in BURFORD [1965].

The accuracy estimates of the computed displacements are best obtained from the covariance matrices of  $\hat{x}^{(1)}$  and  $\hat{x}^{(2)}$ . Applying the covariance law to (2), one obtains

$$\mathbf{C}_{\Delta \hat{x}} = \mathbf{C}_{\hat{x}}^{(1)} + \mathbf{C}_{\hat{x}}^{(2)}. \quad (27.14)$$

This covariance matrix can then be interpreted in terms of *absolute confidence ellipses for displacements* in exactly the same manner as in the case of horizontal positioning (cf §18.3). When absolute confidence ellipses are used, it must be borne in mind that they depend on the choice of constraints imposed. If, for instance, the common origin of the coordinate systems used in both solutions  $\hat{x}^{(1)}$  and  $\hat{x}^{(2)}$  is held fixed, then the confidence ellipse of the origin shrinks to zero, unless otherwise specified.

This situation mirrors that of positioning as discussed in §18.3. Also the absolute ellipses are affected by the indeterminacies of the solution: different particular solutions have different absolute confidence ellipses. An example of the use of absolute confidence ellipses for displacements, when one point  $P_0$  has been assumed fixed, is shown in FIG. 6 [MILLER ET AL., 1969].

As in positioning, it should be considered preferable to work with *relative* rather than absolute *confidence ellipses for the displacements*. These are arrived at using exactly the same procedures as those used in the computation of relative confidence ellipses for positions (see §18.3). The interpretation of one such ellipse pertaining to, say, points  $P_j$ ,  $P_l$  is in terms of confidence limits placed on the relative displacement of  $P_j$  with respect to  $P_l$ , or vice versa. These two equivalent situations are portrayed in FIG. 7. The relative confidence ellipses are not affected by the indeterminacy in translation. The other two indeterminacies, however, have an effect, and thus the constraints that may have to be imposed on rotation and scale must have a geophysical justification if the relative confidence ellipses are to be of any real value.

Using one or the other kind of confidence ellipses, the significance of the derived displacements, from the statistical point of view, can be assessed. The procedure for the assessment is identical with that used in positioning (§18.3) and is not going to

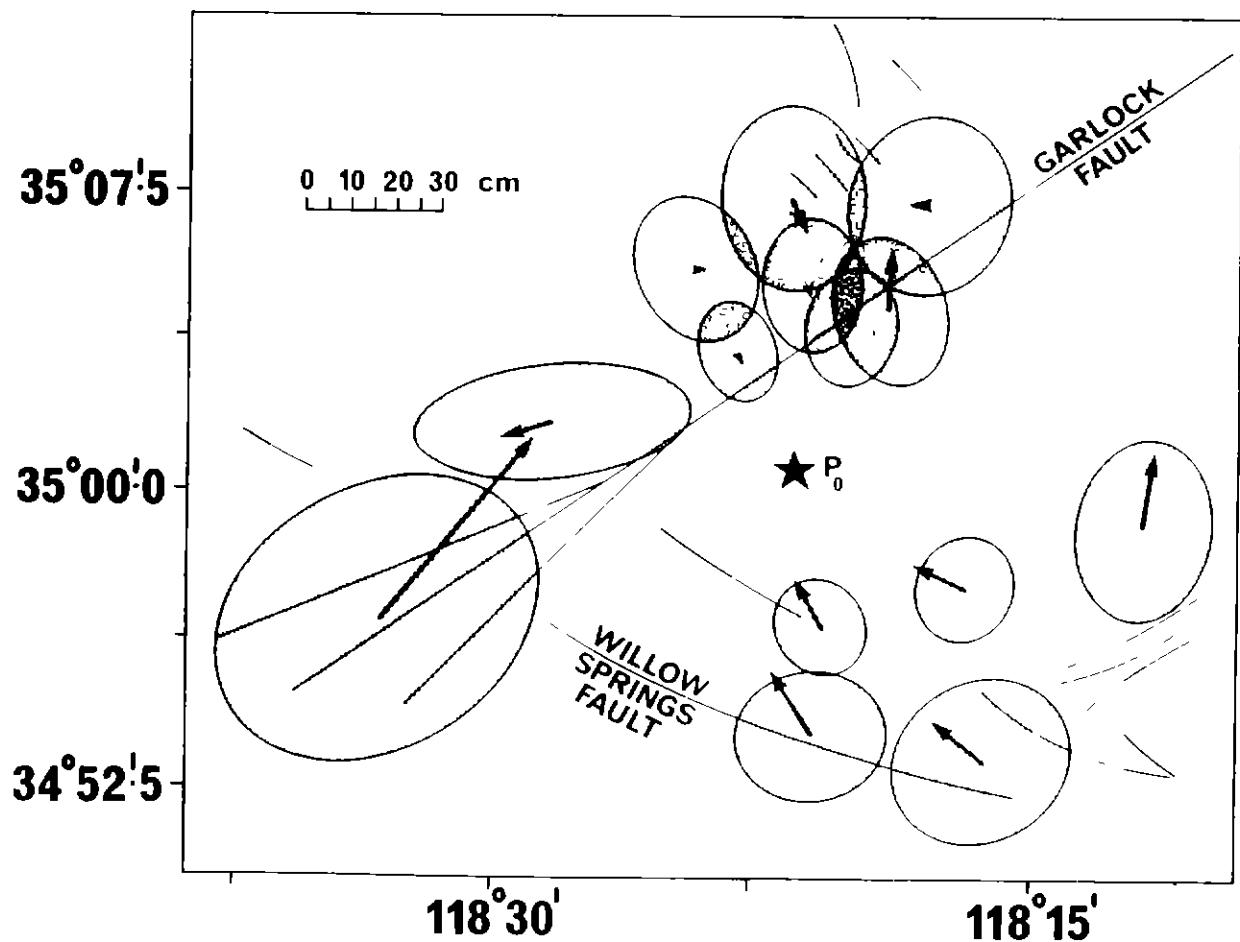


FIG. 27.6. Absolute confidence ellipses of displacements. Point  $P_0$  held fixed.

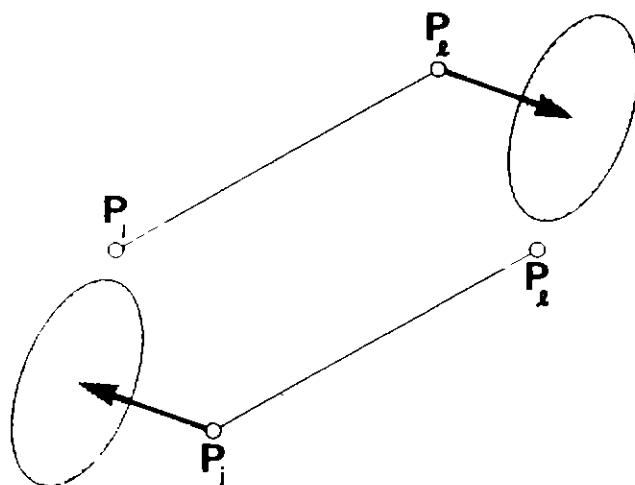


FIG. 27.7. Relative confidence ellipses of displacements.

be repeated here. There is, however, an additional problem here one should be aware of: using the position comparisons, any cross-covariance between  $\mathbf{l}_1$  and  $\mathbf{l}_2$  is usually neglected. This neglect leads to a situation similar to the one already encountered in the context of vertical movements, i.e., to the underestimation of the accuracy of detected displacements, if not to different results altogether. This shortcoming will be discussed at length in the next section.

Under what circumstances should one then use this approach? Clearly, only if there are sound geophysical grounds for imposing the constraints necessary to obtain a solution. Even then one should make sure that the network configuration is favourable so that the improvement to observations through position adjustment ultimately improves the accuracy of the detected displacements. If either of these conditions is not met, then one of the techniques described in the next two sections should be considered instead.

### 27.3. Direct evaluation of horizontal displacements

In the previous section, it was assumed that the rather obvious requirement that  $\mathbf{x}^{(1)}$  and  $\mathbf{x}^{(2)}$  represent positions of the same set of points, i.e., that they be from the same parameter space

$$\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathcal{X}, \quad (27.15)$$

be satisfied. On the other hand,  $\mathbf{l}_1$  and  $\mathbf{l}_2$  may have represented two different sets of observables, i.e., they may have belonged to two entirely different observation spaces  $\mathcal{L}_1$  and  $\mathcal{L}_2$ . To begin with, in this section it will be assumed, for simplicity, that even the observation vectors describe the same observables, i.e., that they belong to the same observation space

$$\mathbf{l}_1, \mathbf{l}_2 \in \mathcal{L}. \quad (27.16)$$

This means that not only the same network of points but also the same observables in that network are considered. Consequently, the two systems of linearized observation equations

$$\mathbf{A}_1 \boldsymbol{\delta}^{(1)} = \mathbf{w}_1, \quad \mathbf{A}_2 \boldsymbol{\delta}^{(2)} = \mathbf{w}_2, \quad (27.17)$$

have the same design matrices; clearly the Jacobian matrices (cf. §12.1) are the same when identical initial values  $\mathbf{x}^{(0)}$  are selected for both epochs  $\tau_1, \tau_2$ . Denoting  $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{A}$ , one can subtract the first equation (17) from the second and get

$$\mathbf{A}(\boldsymbol{\delta}^{(2)} - \boldsymbol{\delta}^{(1)}) = \mathbf{l}_2 - \mathbf{l}_1, \quad (27.18)$$

taking  $\mathbf{f}(\mathbf{x}^{(0)}, \mathbf{l}_1)$  in both cases. Denoting further  $\mathbf{l}_2 - \mathbf{l}_1$  by  $\Delta\mathbf{l}$ , and realizing that  $\boldsymbol{\delta}^{(2)} - \boldsymbol{\delta}^{(1)}$  is nothing less than  $\Delta\mathbf{x}$  (cf. (2)), (18) can be rewritten as the *simple displacement model*

$$\boxed{\mathbf{A} \Delta\mathbf{x} = \Delta\mathbf{l}.} \quad (27.19)$$

An attempt can now be made to obtain the least-squares solution  $\Delta\hat{\mathbf{x}}$ . First, it is necessary to formulate the covariance matrix for  $\Delta\mathbf{l}$ . Using covariance law, one easily obtains

$$\mathbf{C}_{\Delta\mathbf{l}} = \mathbf{C}_{l_1} - 2\mathbf{C}_{l_1 l_2} + \mathbf{C}_{l_2}, \quad (27.20)$$

where  $\mathbf{C}_{l_1 l_2}$  is the cross-covariance for the two sets of observations which, contrary to §27.2, one is now forced to consider. Then the least-squares solution is given by the system of normal equations:

$$\mathbf{A}^T \mathbf{C}_{\Delta\mathbf{l}}^{-1} \mathbf{A} \Delta\hat{\mathbf{x}} = \mathbf{N} \Delta\hat{\mathbf{x}} = \mathbf{A}^T \mathbf{C}_{\Delta\mathbf{l}}^{-1} \Delta\mathbf{l}. \quad (27.21)$$

It is revealing to compare this solution with the one we would get from the same data using the approach described in the previous section. Let us do the comparison systematically:

(a) First, it is left to the reader to show that if  $\mathbf{C}_{l_1 l_2} \neq \mathbf{0}$  or if  $\mathbf{C}_{l_1} \neq \mathbf{C}_{l_2}$  (except when  $\mathbf{C}_{l_2} = k\mathbf{C}_{l_1}$ ,  $k > 0$ ), or if both of these conditions are satisfied, the results obtained by the two techniques will be quite different.

(b) Next, let us assume  $\mathbf{C}_{l_1} = k\mathbf{C}_{l_2}$ , where  $k > 0$ , and  $\mathbf{C}_{l_1 l_2} = \mathbf{0}$ : this is the only case allowing a direct comparison of the presently discussed technique with the one described in §27.2. One gets

$$\mathbf{C}_{\Delta\mathbf{l}} = (k+1)\mathbf{C}_{l_2} = 2\mathbf{C}_{l_1}, \quad (27.22)$$

and the reader can satisfy himself that both techniques yield the same solution; namely,

$$\Delta\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{C}_{l_1}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C}_{l_1}^{-1} \Delta\mathbf{l}. \quad (27.23)$$

Also, the estimated covariance matrix of  $\Delta\hat{\mathbf{x}}$  obtained by both methods is identical, i.e.,

$$\hat{\mathbf{C}}_{\Delta\hat{\mathbf{x}}} = \hat{\sigma}_0^2 (\mathbf{A}^T \mathbf{C}_{l_1}^{-1} \mathbf{A})^{-1}. \quad (27.24)$$

If  $k=1$ , then even the covariance matrices  $\mathbf{C}_{\Delta\hat{x}}$  are identical. Thus, under these rather special circumstances, both techniques are equivalent.

(c) Further, let  $\mathbf{C}_{l_1} = \mathbf{C}_{l_2} = \mathbf{C}_l$  and  $\mathbf{C}_{l_1 l_2} \neq \mathbf{0}$ . The covariance matrix of  $\Delta l$  is now given by (20). Writing

$$\mathbf{C}_{\Delta l}^{-1} = (2\mathbf{C}_l - 2\mathbf{C}_{l_1 l_2})^{-1} = \frac{1}{2}\mathbf{C}_l^{-1} + \delta W, \quad (27.25)$$

the present technique gives

$$(\mathbf{A}^T \mathbf{C}_{\Delta l}^{-1} \mathbf{A} + \mathbf{A}^T \delta W \mathbf{A}) \Delta \hat{x} = (\mathbf{A}^T \mathbf{C}_{\Delta l}^{-1} + \mathbf{A}^T \delta W) \Delta l, \quad (27.26)$$

which is clearly different from the result obtained by position comparison.

(d) Finally, one particular case of (c) deserves special attention because it offers insight into the intrinsic difference of the two approaches: let  $\mathbf{C}_l$  be diagonal, namely,

$$\mathbf{C}_l = \text{diag}(\sigma_{l_i}^2), \quad (27.27)$$

and let  $\mathbf{C}_{l_1 l_2}$  also be diagonal:

$$\mathbf{C}_{l_1 l_2} = \text{diag}(\sigma_{l_i^{(1)} l_i^{(2)}}^2). \quad (27.28)$$

This represents a fairly realistic setup, where statistical dependence is assumed to exist only between the observations of the same observable at the two different epochs. Denoting the correlation coefficients by  $\rho_i$ , one obtains

$$\mathbf{C}_{l_1 l_2} = \text{diag}(\rho_i \sigma_{l_i}^2) = \mathbf{C}_l \text{diag} \rho_i, \quad (27.29)$$

and

$$\mathbf{C}_{\Delta l} = 2(\mathbf{C}_l - \mathbf{C}_l \text{diag} \rho_i) = 2\mathbf{C}_l(I - \text{diag} \rho_i). \quad (27.30)$$

Substituting further the mean value  $\rho$  for the individual  $\rho_i$ , the following equations are obtained (cf. (26.18))

$$\mathbf{C}_{\Delta l} \doteq 2\mathbf{C}_l(1 - \rho), \quad (27.31)$$

and

$$\mathbf{C}_{\Delta l}^{-1} = \frac{1}{2(1 - \rho)} \mathbf{C}_l^{-1}. \quad (27.32)$$

Hence, in this particular case, both techniques yield identical results  $\Delta \hat{x}$ . The covariance matrix  $\mathbf{C}_{\Delta\hat{x}}$  derived from the simple displacement model is, however,  $(1 - \rho)$ -times smaller than that estimated from position comparisons, which is the same as in the case of relevelled segments (cf. §26.3).

When positive statistical dependence is taken into account, it causes the accuracy of the results to increase. Clearly, the simple displacement model is thus mathematically superior to repeated position adjustments because it allows the user to take advantage of the almost certainly present positive statistical dependence between observations of the same observables taken at different epochs. Practically, the assessment of the statistical dependence, compared with its incorporation, is a much more difficult problem deserving considerable research.

The problem of indeterminacy persists, of course, even in the displacement model. However, one additional option appears here: to treat the indeterminacy as a problem with a singularity (cf. §14.5) and use a singular inverse technique [BRUNNER, 1979]. The pseudo-inverse, for example, was shown by POPE AND STEARN [1964] to be equivalent to the approach described in the previous section by (12). No singular inverse technique can make the solution more physically meaningful unless there is a further physical justification for using the technique.

Let us now turn to a more general case when the observations have not been acquired at two sharply defined epochs  $\tau_1, \tau_2$  but rather collected over a period of time. It is then meaningless to speak of  $I_1, I_2$ , and the model must account for each individual observation  $I_j$  made at time  $\tau_j$ . The only way of designing such a model is to hypothesize certain properties of the displacement field. To show how this technique works, let us begin by postulating that the horizontal movements are linear in time. Then the relation between the horizontal velocities  $\dot{z}^*$  postulated to be constant in time, and the displacements  $\Delta z^*$ , can be formulated as follows: for the point  $P_j$  we have

$$\dot{z}_j^* = \dot{x}_j + i \dot{y}_j, \quad (27.33)$$

and the sought relation is evidently

$$\boxed{\Delta z^*(\Delta\tau) = \dot{z}^* \cdot \Delta\tau.} \quad (27.34)$$

As to the velocities, they can be linked with changes in the observables. The equations for these relations are easily obtained from the observation equations used in positioning. Taking, for instance, the linearized equation for ‘observed’ chord distance  $I_{jk}$  (18.13), one has

$$I_{jk} = A_{jk}(l) [\delta x_j, \delta y_j, \delta x_k, \delta y_k]^T + I_{jk}^{(0)} - r_{jk}^l, \quad (27.35)$$

where  $A_{jk}(l)$  is a function of the azimuth of the line. When temporal changes only in the observed distance are considered, differentiation with respect to time yields

$$\dot{I}_{jk} = A_{jk}(l) [\delta \dot{x}_j, \delta \dot{y}_j, \delta \dot{x}_k, \delta \dot{y}_k]^T = A_{jk}(l) [\dot{x}_j, \dot{y}_j, \dot{x}_k, \dot{y}_k]^T. \quad (27.36)$$

Realizing that postulated linear movements imply linear changes in observables, the observation equation for the temporal change in observed distances is

$$\frac{I_{jk}(\tau_2) - I_{jk}(\tau_1)}{\tau_2 - \tau_1} = A_{jk}(l) [\dot{x}_j, \dot{y}_j, \dot{x}_k, \dot{y}_k]^T = \text{re } A_{jk}^*(l) [\dot{z}_j^*, \dot{z}_k^*]^T, \quad (27.37)$$

where  $I_{jk}(\tau_1), I_{jk}(\tau_2)$  are the ‘observed’ values of the chord distance  $I_{jk}$  at epochs  $\tau_1, \tau_2$ , and  $A_{jk}^*(l)$  is the complex form of  $A_{jk}(l)$ , which the reader can easily derive.

Analogous equations can be written for angles, directions, and azimuths. The whole system of observation equations may then be assembled to create the constant

velocity displacement model

$$\boxed{\mathbf{A}\dot{\mathbf{x}} = \text{re } \mathbf{A}^* \dot{\mathbf{z}}^* = \mathbf{i}.} \quad (27.38)$$

It should be noted that multiplication by  $\Delta\tau$  reduces this model to the form of the simple displacement model (19). The only difference is that here each component of the vector of observed velocities may be referred to time intervals of different lengths. The displacements  $\Delta\hat{z}^*$  are then evaluated in exactly the same way as in the case of the simple displacement model.

When some observables have been reobserved more than once, not only horizontal velocities but also horizontal accelerations of some points may be determined. Following a similar reasoning as above, the *constant acceleration displacement model*

$$\boxed{\mathbf{A}\ddot{\mathbf{x}} = \text{re } \mathbf{A}^* \ddot{\mathbf{z}}^* = \ddot{\mathbf{i}}.} \quad (27.39)$$

can be written. This equation, together with (38), gives the complete information on the movement which may be characterized by accelerations constant in time. Expressing the second-order derivative through finite differences [NORRIE AND DE VRIES, 1978], the  $j$ th quasi-observation becomes

$$\ddot{i}_j = \frac{i_j(\tau_3) - 2i_j(\tau_2) + i_j(\tau_1)}{(\tau_3 - \tau_2)(\tau_2 - \tau_1)}. \quad (27.40)$$

Here it is tacitly assumed that the observations were acquired at consecutive epochs, i.e.,

$$\tau_1 < \tau_2 < \tau_3. \quad (27.41)$$

The constant velocity and constant acceleration models may be considered as two special cases of *temporally constrained displacement models*. An alternative family of models that can be used for scattered observations are *spatially constrained displacement models*. A wide selection of these models is discussed in the literature, see, e.g., POPE AND STEARN [1964]. Any such selection should, ideally, stem from the knowledge or at least from the postulation of the physical laws that govern the movements. As an example, the motion may be postulated to consist of a pure slip in a constant direction, usually parallel to a known active fault. Specifics of this *slip displacement model* may be found in, e.g., FRANK [1966]. Another example of a spatially constrained model is the one that prescribes the magnitude of the displacement to change with location according to a postulated law. Such a *constrained magnitude displacement model* may be encountered when interpolate coseismic movements are investigated: there the reduction of the displacement magnitude with distance from the fault is a function of the depth of the epicentre. An illustration of such a displacement behaviour for the earlier cited Tango earthquake, according to KASAHARA [1957], is shown in FIG. 8. All these models are merely constrained models in the sense discussed in §14.5. Their mathematical development is therefore left to the reader.

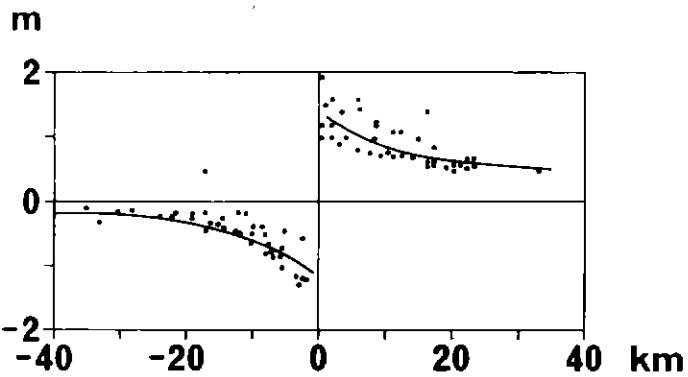


FIG. 27.8. Decrease in the magnitude of horizontal displacement with distance from fault.

The last models that need be discussed in this section are the sub-family of the spatially constrained models, the *spatially continuous displacement models*. The idea behind these models parallels that of the areal models for vertical displacements (cf. §26.4). Usually, these models seek the horizontal displacements in terms of series of algebraic functions of positions, i.e. [KASAHARA AND SUGIMURA, 1964; WHITTEN, 1967],

$$\Delta z^*(z^*) = \sum_{j=0}^m c_j^* z^{*j}. \quad (27.42)$$

The mechanics of obtaining the coefficients  $c_j^*$  is the same as those used in §26.4 and will not be enlarged on here. It is, however, interesting to have a look at some results obtained by means of this model: FIG. 9 shows horizontal displacements in Imperial Valley, California, as computed by SNAY AND GERGEN [1978]. It should be noted that here the authors assumed spatial continuity only within the areas delimited by the two active faults.

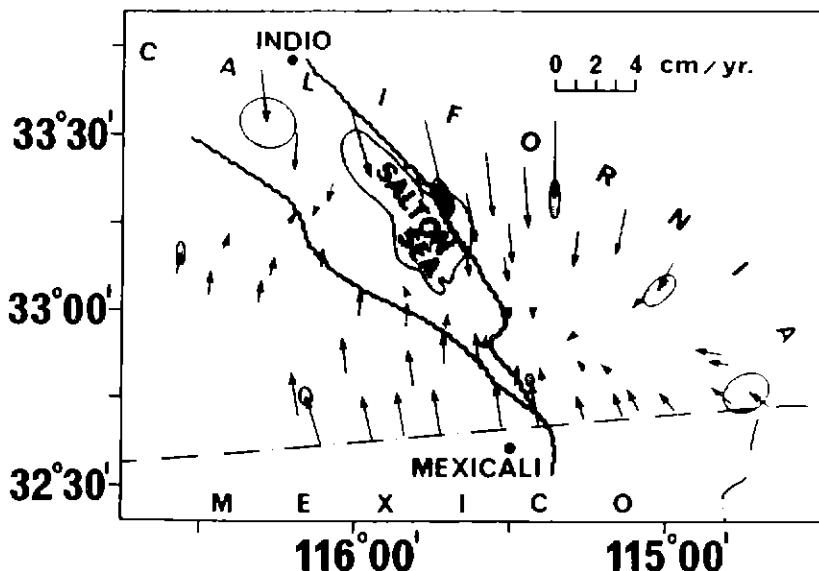


FIG. 27.9. Mean annual horizontal displacements in Imperial Valley, California, computed from data covering the period 1941 to 1975.

The common disadvantage of the direct approach to displacements, compared with the repeated position approach, is that observations which have not been repeated at a later epoch cannot be accommodated within the model. On the other hand, the estimation of errors with the displacement models should be considered superior to that with the position comparison, for reasons explained above. Also, computationally the direct approach is preferable because fewer unknown parameters are sought; there is no need to evaluate positions. None of the constrained models is again free of indeterminacy, which persists in one form or another. The presence of the indeterminacy problem is clearly symptomatic of the fact that the displacements are not the most natural quantities to look for; some more natural alternatives to displacements will be shown in the last section.

#### 27.4. Strain, shear, and other models

An obvious way of alleviating the indeterminacy is to look at other quantities that describe horizontal movements and are more directly linked with the observables. The most widely used quantity, in this respect, is the strain tensor (cf. §25.2). To show how it is used in horizontal movement studies, let us begin by rewriting the two-dimensional strain tensor (25.26) in the mapping coordinate system  $(x, y)^M$ . This operation yields

$$\boldsymbol{\epsilon}' = \begin{bmatrix} e_{xx} & e_{xy} \\ e_{yx} & e_{yy} \end{bmatrix} = \begin{bmatrix} \frac{\partial v_x}{\partial x} & \frac{1}{2} \left( \frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x} \right) \\ \frac{1}{2} \left( \frac{\partial v_y}{\partial x} + \frac{\partial v_x}{\partial y} \right) & \frac{\partial v_y}{\partial y} \end{bmatrix}. \quad (27.43)$$

Denoting the two-dimensional  $\nabla$  operator by  $\nabla'$ , one can also write

$$\boldsymbol{\epsilon}' = \frac{1}{2} \left[ \nabla' \bar{v}^T + (\nabla' \bar{v}^T)^T \right], \quad (27.44)$$

where  $\bar{v} = (v_x, v_y)^T$  is the displacement vector as defined by (3).

It is interesting to note that the strain tensor may also be considered as the symmetrical part of the Jacobian matrix of transformation from the position vector space to displacement vector space, or, alternatively, as the measure of the *symmetrical deformation*. More specifically, for small deformations, one can write [JAEGER, 1971]

$$\bar{v} \doteq (\Omega' + \boldsymbol{\epsilon}') \bar{r} + \bar{v}_0, \quad (27.45)$$

where  $\Omega'$  is the antisymmetrical part of the Jacobian matrix and  $\bar{v}_0$  is an arbitrary translation. The symmetrical part of the displacement, i.e.,  $\boldsymbol{\epsilon}' \bar{r}$ , describes that part of the deformation that does not allow for any rotations. This is the part we will now concentrate on. The antisymmetrical part will be dealt with later. Other decompositions of the *displacement gradient matrix*  $\Omega' + \boldsymbol{\epsilon}'$  are used. In some applications, for instance, the decomposition into conformal and anticonformal parts is advantageous (e.g., SCHNEIDER [1982]).

Because of the differentiation of  $\bar{v}$  involved in the evaluation of  $\epsilon'$ , the translation indeterminacy that plagues  $\bar{v}$  (cf. (7)) disappears, i.e.,  $\epsilon'$  is insensitive to translations. The indeterminacy in rotation is characterized by the term  $z_p^{*(1)}(e^{-i\Delta\psi} - 1)$  (cf. (10)). In the present context, it is preferable to express the rotation by means of matrices:

$$(\mathbf{R}_t(\Delta\psi) - \mathbf{I})\bar{r} = \mathbf{S}(\Delta\psi)\bar{r}, \quad (27.46)$$

where we can think of the  $\mathbf{S}(\Delta\psi)$  matrix as a one-parametric rotational displacement operator; note that  $\mathbf{S}(0)=\mathbf{0}$ . In this notation, the complete displacement  $\bar{v}_c$  of a point can be written as

$$\bar{v}_c = \bar{v}_p + \mathbf{S}(\Delta\psi)\bar{r} + \bar{v}_0, \quad (27.47)$$

where  $\bar{v}_p$  is a particular value of the displacement vector. Substitution of the complete solution into (44) for strain tensor, yields

$$\epsilon'_c = \frac{1}{2} \left\{ \nabla' (\bar{v}_p + \mathbf{S}\bar{r})^T + [\nabla' (\bar{v}_p + \mathbf{S}\bar{r})^T]^T \right\}. \quad (27.48)$$

Application of the rules of vector differentiation (see §3.2) gives

$$\nabla' (\bar{v}_p + \mathbf{S}\bar{r})^T = \nabla' \bar{v}_p^T + \mathbf{S}^T, \quad (27.49)$$

and, consequently,

$$\epsilon'_c = \frac{1}{2} \left[ \nabla' \bar{v}_p^T + (\nabla' \bar{v}_p^T)^T \right] + \frac{1}{2} (\mathbf{S} + \mathbf{S}^T). \quad (27.50)$$

A brief calculation shows that

$$\frac{1}{2} [\mathbf{S}(\Delta\psi) + \mathbf{S}^T(\Delta\psi)] = (\cos \Delta\psi - 1)\mathbf{I}, \quad (27.51)$$

so that the complete solution for the strain tensor reads

$$\epsilon'_c = \frac{1}{2} \left[ \nabla' \bar{v}_p^T + (\nabla' \bar{v}_p^T)^T \right] - (1 - \cos \Delta\psi)\mathbf{I}. \quad (27.52)$$

Equation (52) shows that the indeterminacy in rotation of displacements transforms into the indeterminacy in the diagonal components of the strain tensor, i.e., into indeterminacy in strain in  $x$  and  $y$  directions. In reality, when there is a rotation around the origin of the coordinate system, it is quite small so that the uncertainty may be written as

$$-(1 - \cos \Delta\psi) \doteq \frac{1}{2} \Delta\psi^2. \quad (27.53)$$

It thus represents a second-order effect, compared with the first-order effect, on displacements characterized by the presence of  $\sin \Delta\psi$ . For example, an unrecognized rotation as large as one minute of arc causes an error smaller than  $5 \times 10^{-8}$  in the solution for strain in the  $x$  and  $y$  directions. Hence, in addition to being invariant with respect to translations, the strain tensor is also near-invariant with respect to (small) rotations, which is what makes the tensor so attractive a choice for describing horizontal movements.

Turning finally to the scale indeterminacy, one discovers that it affects the strain tensor the same way it affects the displacements—see eqn. (11). The complete solution of the strain tensor for small  $\Delta\psi$  then reads

$$\epsilon'_c \doteq \frac{1}{2} \left[ \nabla' \bar{v}_p^T + (\nabla' \bar{v}_p^T)^T + \Delta\psi^2 I \right] (1 + \Delta\kappa). \quad (27.54)$$

The additional value of this equation for studying the effect of systematic errors in azimuths (average error in  $\alpha = \Delta\psi$ ) and distances (average error in  $\Delta r = \Delta\kappa$ ) on the evaluated strain tensor should be self-evident.

How does one obtain the strain tensor from observed distances, angles, and azimuths? The most natural way is to first evaluate the displacement vectors  $\bar{v}_p$  for all the points in the network, using one of the techniques described in §27.3. Then the partial derivatives composing the strain tensor can be evaluated numerically from the displacements of adjacent points. The configurations of adjacent points for this task may be selected in a variety of modes. The natural selection is the one using the basic configurations of the network, which are usually triangles. Taking one triangle  $P_i, P_j, P_k$  as the basic configuration, we are faced with the situation depicted in FIG. 10.

Clearly, even when the basic configuration has been selected, the four partial derivatives needed in  $\epsilon'$  may still be evaluated in several different ways. The most straightforward technique appears to be the linear interpolation amongst  $P_i, P_j, P_k$  the results of which are then associated with the centroid  $P_{ijk}$ . This approach was first proposed by TERADA AND MIYABE [1929] and may be described as follows. Let us begin by writing the equation for the plane representing the  $x$ -component of the displacement vectors in the triangle as (see §3.3)

$$v_x(x, y) = \frac{\partial v_x}{\partial x} x + \frac{\partial v_x}{\partial y} y + \text{const.} \quad (27.55)$$

Substituting the specific values of  $x, y, v_x(x, y)$  for the three points, three linear, simultaneous equations for  $\partial v_x / \partial x, \partial v_x / \partial y$  and the constant term are obtained. A similar system is then written for  $v_y$ . These two linear systems are solved for all four

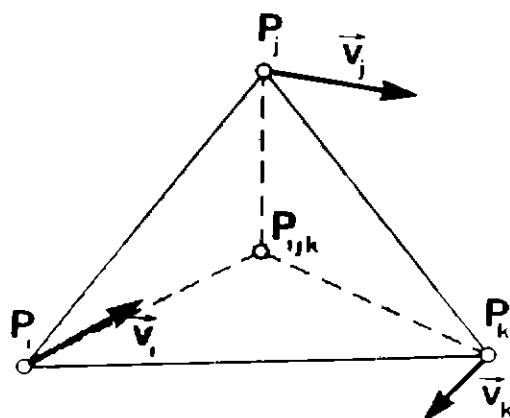


FIG. 27.10. Finite-element-like basic strain configuration.

partial derivatives and the two constant terms (which are of no consequence here).

Another choice of the basic configuration is to use all points connected to the point of interest  $P_i$  by observations of one kind or another. This configuration is shown in FIG. 11. Here, the characteristic values of the partial derivatives of displacements are obtained through the least-squares estimation [VANÍČEK ET AL., 1981] and associated with  $P_i$ . The difference between the two approaches is analogous to the difference between the finite element and finite difference approaches to the numerical solution of a boundary value problem.

For other possible techniques, the reader is referred to, e.g., TSUBOI [1930]. It should be noted that the covariance matrix  $C_\epsilon$  pertaining to the four elements of the strain tensor can be evaluated through the covariance law. The derivation is left to the reader.

Once the strain tensor has been derived, it can supply, of course, all the information about the average state of strain in the triangle  $P_i, P_j, P_k$  or at a point. This information may then be displayed in the form of a conic corresponding to the strain tensor or its pedal curve, much the same way as was done in §16.3 for map distortions (Tissot's indicatrix) or in §25.2 for the tidal strain. This technique is similar to that of constructing confidence ellipses from the two by two submatrices of the weight matrix (cf. §18.3) and will not be repeated here. On the other hand, an interpretation of such a *strain conic* is needed.

FIG. 12 shows a possible strain conic, an ellipse: the axes are directed along the eigenvectors of the strain tensor, and maximum and minimum strains,  $\epsilon_1, \epsilon_2$ , i.e., the eigenvalues of the strain tensor (and not  $\epsilon_1^{-2}, \epsilon_2^{-2}$ ), are shown as half-lengths of the axes. Unlike covariance matrices or their diagonal submatrices, strain tensors may have either positive or negative eigenvalues. A positive value signifies *extension* in the direction of the corresponding axis; a negative value shows *contraction*, see, e.g., DERMANIS AND LIVIERATOS [1983]. Clearly, the strain conic is an ellipse only when both eigenvalues are positive. When one is positive and the other is negative, the conic is a two-branch hyperbola. In this case, authors often plot a rosette similar to

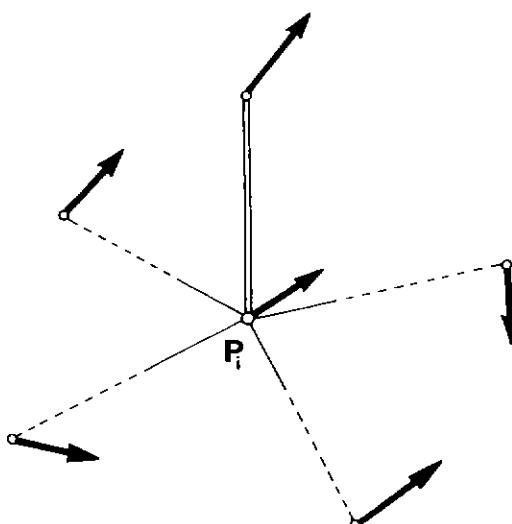


FIG. 27.11. Finite-difference-like basic strain configuration.

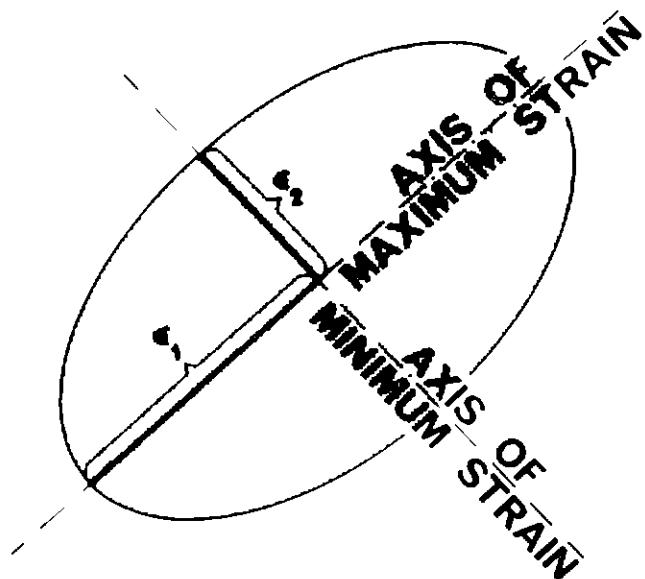


FIG. 27.12. Strain ellipse.

the one shown in FIG. 14. When one eigenvalue is equal to zero, the conic becomes a segment; and when both  $\epsilon_1$  and  $\epsilon_2$  are negative, we have an imaginary ellipse. Most authors, however, do not distinguish among the different ellipses and hyperbolae and plot all the cases, when  $\epsilon_1\epsilon_2 \neq 0$ , as ellipses. It is also common practice that instead of plotting the whole strain conic, only the *principal axes of strain* are plotted. To show the character of strain along these axes, different lines (e.g., solid lines for extension, dashed for contraction) are used. FIG. 13 shows a strain pattern computed by TSUBOI [1932] using the same data as were used for constructing FIG. 2. The reader is advised to compare the two kinds of displays.

A very popular way of portraying horizontal strain is by means of *shear strains*. These quantities are defined as

$$\gamma_1 = \epsilon_{xx} - \epsilon_{yy}, \quad \gamma_2 = 2\epsilon_{xy}, \quad (27.56)$$

and represent shear in directions inclined to the  $x$  and  $y$  axes by an angle of  $\frac{1}{4}\pi$ . Interestingly, shear strains are completely insensitive to rotations (cf. (54)). Shear strain vanishes in the direction of the principal strain axes and reaches the maximum, called *total shear*,

$$\gamma_m = \sqrt{\gamma_1^2 + \gamma_2^2}, \quad (27.57)$$

in directions inclined to the principal axes by an angle of  $\frac{1}{4}\pi$ . Evidently, shear vanishes if the strain conic is a circle. The systematically plotted variation of shear with azimuth, known as *shear rosette*, is shown in FIG. 14. More details on shear may be found, for instance, in JAEGER [1971]. Radial expansion

$$\rho = \frac{1}{2}(\epsilon_1 + \epsilon_2), \quad (27.58)$$

called *dilation*, can also be computed to advantage. The interested reader is advised to look up illustrations of how all these quantities are interpreted geophysically in, e.g., RIKITAKE [1976].

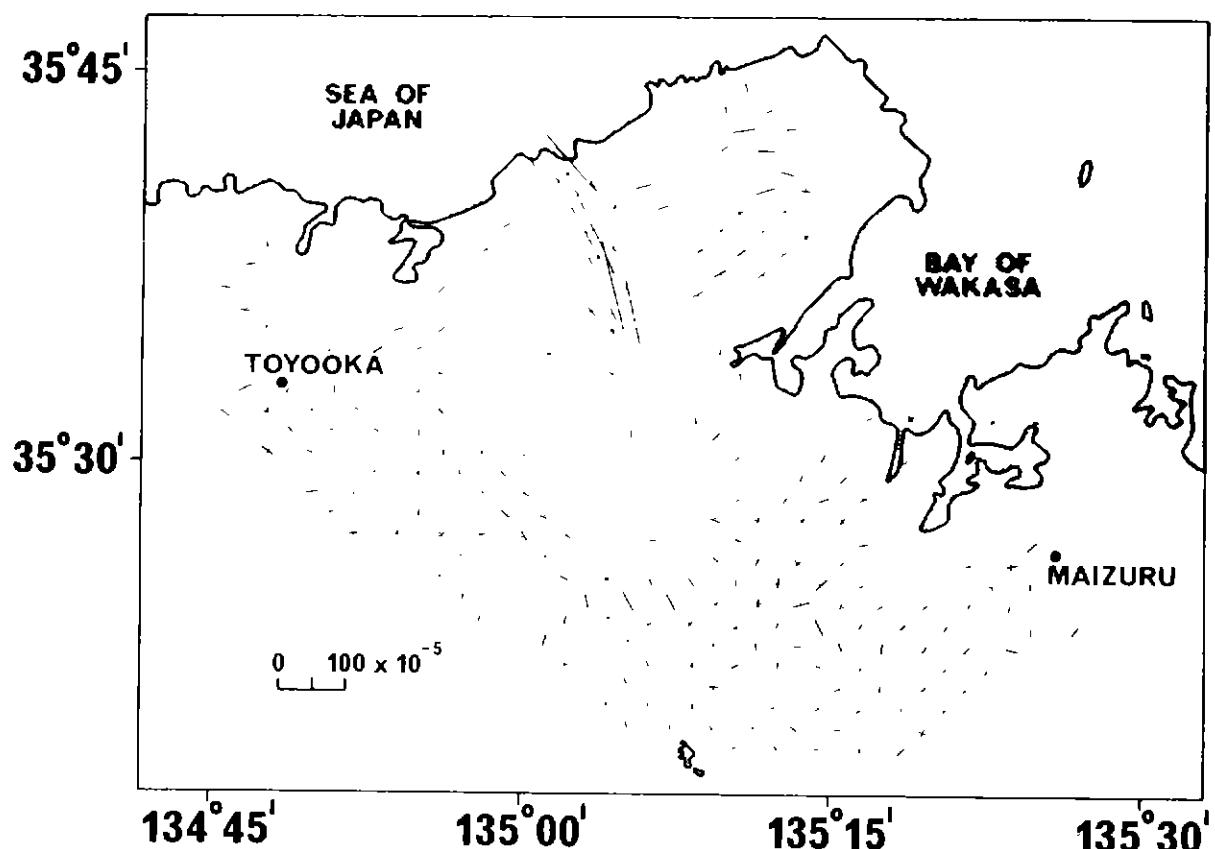


FIG. 27.13. Principal axes of strain. Solid lines show extension, dashed lines contraction.

Clearly, all the above treated guises of strain can be derived from the strain tensor, and thus they all describe the symmetrical deformation. As such, they are all insensitive to translations and almost so to rotations. The situation is different for the  $\Omega'$  which measures the *antisymmetrical deformation*. From (45) we can write

$$\Omega' = \begin{bmatrix} 0 & \frac{1}{2} \left( \frac{\partial v_x}{\partial y} - \frac{\partial v_y}{\partial x} \right) \\ \frac{1}{2} \left( \frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right) & 0 \end{bmatrix} = \begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix}, \quad (27.59)$$

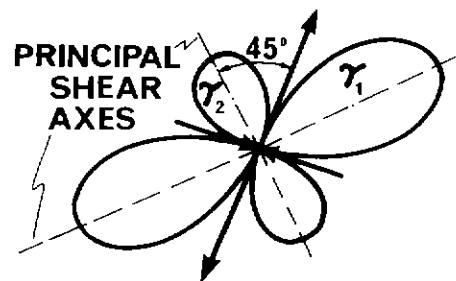


FIG. 27.14. Shear rosette.

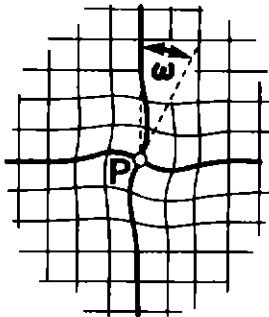


FIG. 27.15. Average differential rotation.

where  $\omega$  is known as the *average differential rotation* of the neighbourhood of  $P$  with respect to the coordinate system—see FIG. 15. It can be shown, by means similar to those used in deriving (52), that the complete solution for  $\Omega'$  reads

$$\begin{aligned} \Omega'_c &= \begin{bmatrix} 0 & \frac{1}{2} \left( \frac{\partial v_x}{\partial y} - \frac{\partial v_y}{\partial x} \right) + \sin \Delta\psi \\ \frac{1}{2} \left( \frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right) - \sin \Delta\psi & 0 \end{bmatrix} \\ &\doteq \begin{bmatrix} 0 & \omega + \Delta\psi \\ -\omega - \Delta\psi & 0 \end{bmatrix}. \end{aligned} \quad (27.60)$$

Thus the average differential rotation suffers from the indeterminacy in rotation much the same way as do the displacement vectors.

It should be mentioned that all the strain related quantities can be derived directly from observations without having to evaluate the displacements first. Practical formulae using only observed changes in angles have been derived by FRANK [1966]. A variation of Frank's approach has been proposed by BIBBY [1975].

To conclude, let us state that it is highly desirable to use a combination of horizontal and vertical information to arrive at a three-dimensional description of the deformations. Such a three-dimensional model is more easily converted to geophysically meaningful quantities such as stresses and forces—see, e.g., NYLAND [1977]. Unfortunately, however, the available data are usually either vertical or horizontal, but very rarely both.

## PART VI

### REFERENCES

- ALTERMAN, Z., H. JAROSCH AND C.L. PEKERIS (1961). Propagation of Rayleigh waves in the earth. *Geophys. J. Roy. Astronom. Soc.* 4, p. 219.
- BAKER, T.F. AND G.W. LENNON (1976). Spatial coherency and tidal tilts. *Proc. 7th International Symposium on Earth Tides*, Ed. C. Szádeczky-Kardoss. Hungarian Academy of Sciences, Sopron, Hungary, September, 1973. E. Schweizerbart'sche Verlagsbuchhandlung, pp. 479-493.
- BEAUMONT, C. (1976). Personal communication. Dalhousie University, Halifax, Canada.
- BEAUMONT, C. AND A. LAMBERT (1972). Coastal structure from surface load tilts using a finite element model. *Geophys. J. Roy. Astronom. Soc.* 29, pp. 203-226.
- BENDER, P.L., J.E. FALLER, J. LEVINE (AND OTHERS) (1979). Possible high-mobility LAGEOS ranging station. *Tectonophysics* 52, pp. 69-73.
- BIBBY, H.M. (1975). Crustal strain from triangulation in Marlborough, New Zealand. *Tectonophysics* 29, pp. 529-540.
- BOMFORD, G. (1971). *Geodesy*. 3rd ed., Oxford University Press.
- BOWER, D.R. (1970). Some numerical results in the determination of the indirect effect. *Proc. 6th International Symposium on Earth Tides*, Ed. R. Dejaiffe. IUGG and IAG, Strasbourg, France, September, 1969. Observatoire Royal de Belgique, pp. 106-112.
- BOWER, D.R. (1973). A sensitive water level tiltmeter. *Philos. Trans. Roy. Soc. London Ser. A* 274 (1239), pp. 223-226.
- BRAGARD, L. (1980). Personal communication. University of Liege, Belgium.
- BRUNNER, F.K. (1979). On the analysis of geodetic networks for the determination of the incremental strain tensor. *Surv. Rev.* XXV (192), pp. 56-67.
- BURFORD, R.O. (1965). Strain analysis across the San Andreas fault and Coast Ranges of California. *Proc. 2nd Symposium of the Commission on Recent Crustal Movements*, IAG and IUGG, Aulanko, Finland, August.
- BURFORD, R.O., R.D. NASON AND P.W. MARSH (1978). Studies of fault creep in central California. *Earthquake Inf. Bull.* 10 (5), pp. 174-181.
- CABANISS, G.H. (1978). The measurement of long period and secular deformation with deep borehole tiltmeters. *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference, An International Symposium on the Applications of Geodesy to Geodynamics*, Ed. I.I. Mueller. IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280, The Ohio State University, Columbus, U.S.A., pp. 165-169.
- CASTLE, R.O., J.N. ALT, J.C. SAVAGE AND E.I. BALAZS (1974). Elevation changes preceding the San Fernando earthquake of February 9, 1971. *Geology* 2, pp. 61-66.
- CHI, S.C., R.E. REILINGER, L.D. BROWN AND G.A. JURKOWSKI (1980). Geodetic leveling and crustal movement in the U.S., Part II, Non-tectonic influences. Paper presented at the 1980 Spring Meeting of the American Geophysical Union, May 22-27, Toronto, Canada.
- CHRZANOWSKI, A. (1980). Personal communication. Department of Surveying Engineering, University of New Brunswick, Fredericton, Canada.
- CONDON, E.U. AND H. ODISHAW (EDS.) (1967). *Handbook of Physics*. 2nd ed., McGraw-Hill.
- COORDINATING COMMITTEE ON GREAT LAKES BASIC HYDRAULIC AND HYDROLOGIC DATA, THE (1977).

- Apparent vertical movement over the Great Lakes U.S. Army Corps of Engineers, Detroit District, U.S.A.
- CURRIE, R.G. (1975). Period,  $Q_p$ , and amplitude of the pole tide. *Geophys. J. Roy. Astronom. Soc.* 43, pp. 73-86.
- DERMANIS, A. AND E. LIVIERATOS (1983). Applications of deformation analysis in geodesy and geodynamics. *Reviews of Geophysics and Space Physics* 21(1), pp. 41-50.
- DOODSON, A.T. (1922). The harmonic development of the tide-generating potential. *Proc. Roy. Soc. London Ser. A* 100, pp. 305-329.
- EDGE, R.C.A. (1959). Some considerations arising from the results of the second and third geodetic levellings of England and Wales. *Bull. Géod.* 52, pp. 28-36.
- FARRELL, W.E. (1972). Deformation of the earth by surface loads. *Rev. Geophys. and Space Phys.* 10 (3), pp. 761-797.
- FRANK, F.C. (1966). Deduction of earth strains from survey data. *Bull. Seismol. Soc. Amer.* 56 (1), pp. 35-42.
- GOAD, C.C. (1979). Gravimetric tidal loading computed from integrated Green's functions. NOAA Technical Memorandum NOS NGS-22, National Geodetic Survey, Rockville, U.S.A.
- GOODKIND, J.M. (1978). High precision tide spectroscopy. *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference, An International Symposium on the Applications of Geodesy to Geodynamics*, Ed. I.I. Mueller. IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280, The Ohio State University, Columbus, U.S.A., pp. 309-311.
- HARKRIDER, D.G. (1970). Surface waves in multilayered elastic media 2, higher mode spectra and spectral ratios from point sources in plane layered earth models. *Bull. Seismol. Soc. Amer.* 60, p. 1937.
- HOLDAHL, S.R. AND R.L. HARDY (1979). Solvability and multiquadric analysis as applied to investigations of vertical crustal movements. *Tectonophysics* 52, pp. 139-155.
- HOLDAHL, S.R. AND N.L. MORRISON (1974). Regional investigations of vertical crustal movements in the U.S. using precise levellings and mareograph data. *Tectonophysics* 23, pp. 373-390.
- HOLLAND, G.L. AND T.S. MURTY (1970). On the pole tide and related Chandler oscillations. *Report on the Symposium on Coastal Geodesy*, Ed. R. Sigl. IUGG and IAG, Munich, Germany, July. Institut für Angewandte Geodäsie, pp. 369-389.
- INTERNATIONAL ASTRONOMICAL UNION (1977). *Proceedings of the Sixteenth General Assembly*. Ed. A. Muller, A. Jappel. IAU, Grenoble, 1976. Trans. of the IAU, Vol. XVIB, D. Reidel Publishing.
- JACHENS, R.C. (1978). The gravity method and interpretive techniques for detecting vertical crustal movements. *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference, An International Symposium on the Applications of Geodesy to Geodynamics*, Ed. I.I. Mueller. IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280, The Ohio State University, Columbus, U.S.A., pp. 153-155.
- JACHENS, R.C. (1979). Personal communication. U.S. Geological Survey, Menlo Park, U.S.A.
- JAEGER, J.C. (1971). *Elasticity, Fracture, and Flow: With Engineering and Geological Applications*. 3rd ed., Halsted Press.
- KAÄRIÄINEN, E. (1953). On the recent uplift of the earth's crust in Finland. Publication of the Finnish Geodetic Institute No. 42, Helsinki, Finland.
- KASAHARA, K. (1957). The nature of seismic origins as inferred from seismological and geodetic observations (I). *Bulletin of the Earthquake Research Institute, University of Tokyo, Japan*, Vol. 35, pp. 511-530.
- KASAHARA, K. AND A. SUGIMURA (1964). Horizontal secular deformation of land deduced from triangulation data, I. Land deformation in central Japan. *Bulletin of the Earthquake Research Institute, University of Tokyo, Japan*, Vol. 42, pp. 479-490.
- KORHONEN, J. (1961). Adjustment of levellings in region of slow vertical movement. *Ann. Acad. Sci. Fennicae. Ser. AIII, Geologica-Geographica*, pp. 128-142.
- LAMBECK, K. (1980). *The Earth's Variable Rotation: Geophysical Causes and Consequences*. Cambridge University Press.
- LAMBERT, A. AND C. BEAUMONT (1977). Nano variations in gravity due to seasonal groundwater movements: Implications for the gravitational detection of tectonic movements. *J. Geophys. Res.* 82, pp. 297-306.

- LAMBERT, A. AND P. VANIČEK (1979). Contemporary crustal movement in Canada. *Canad. J. Earth Sci.*, 16 (3), part 2, pp. 647–668.
- LAMBERT, W.D. AND F.W. DARLING (1936). Tables for determining the form of the geoid and the indirect effect on geodesy. U.S. Coast and Geodetic Survey Special Publication 199, Washington, D.C., U.S.A.
- LENNON, G.W. AND P. VANIČEK (1970). Calibration tests and the comparative performance of horizontal pendulums at a single station. *Proc. 6th International Symposium on Earth Tides*, Ed. R. Dejaiffe, IUGG and IAG, Strasbourg, France, September, 1969. Observatoire Royal de Belgique, pp 183–193.
- LEVINE, J. (1978). Multiple wavelength geodesy *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference, An International Symposium on the Applications of Geodesy to Geodynamics*, Ed. I.I. Mueller, IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280, The Ohio State University, Columbus, U.S.A., pp. 99–102.
- LONGMAN, I.M. (1962). A Green's function for determining the deformation of the earth under surface mass loads I. *J. Geophys. Res.* 67 (2), pp. 845–850.
- LONGMAN, I.M. (1963). A Green's function for determining the deformation of the earth under surface mass loads II. *J. Geophys. Res.* 68 (2), pp. 485–496.
- LOVE, A.E.H. (1911). *Some Problems of Geodynamics*. Dover reprint, 1967.
- LOVE, A.E.H. (1927). *A Treatise on the Mathematical Theory of Elasticity*. 4th ed., Dover reprint, 1944.
- MEADE, B.K. (1973). Report of the sub-commission on recent crustal movements in North America. In: "Reports on Geodetic Measurements of Crustal Movement, 1906–1971", Paper No. 65, U.S. Department of Commerce, Rockville, U.S.A.
- MEADE, B.K. AND J.B. SMALL (1966). Current and recent movement on the San Andreas fault. *Geology of Northern California*, Bulletin 190, Division of Mines and Geology, pp. 385–391.
- MELCHIOR, P. (1978). *The Tides of the Planet Earth*. Pergamon.
- MILLER, R.W., A.J. POPE, H.S. STETTNER AND J.L. DAVID (1969). Crustal movement investigations—triangulation, Taft-Mojave area, California, supplement. U.S. Coast and Geodetic Survey Operational Data Report C & GS Dr-6, Rockville, U.S.A.
- MUNK, W.M. AND G.F.K. MACDONALD (1960). *The Rotation of the Earth*. Cambridge University Press.
- NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (1979). Application of space technology to crustal dynamics and earthquake research. NASA Technical Paper 1464, Washington, D.C., U.S.A.
- NORRIE, D.H. AND G. DE VRIES (1978). *An Introduction to Finite Element Analysis*. Academic Press.
- NUR, A. AND G. MAVKO (1974). Postseismic viscoelastic rebound. *Science* 183, pp. 204–206.
- NYLAND, E. (1977). Repeated geodetic surveys as experiments in geophysics. *Canad. Surv.* 31 (4), pp. 347–360.
- OZAWA, I. (1961). On the observations of the earth tide by means of extensometers in horizontal components. Disaster Prevention Research Institute Report 46, Kyoto University, Japan.
- POPE, A.J. AND J.L. STEARN (1964). Matrix algebra applied to horizontal earth movement analysis Paper presented at the 45th Annual Meeting of the American Geophysical Union, Washington, D.C., April.
- RIKITAKE, T. (1976). *Earthquake Prediction*. Elsevier.
- SAVAGE, J.C. (1978). Strain patterns and strain accumulation along plate margins. *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference, An International Symposium on the Applications of Geodesy to Geodynamics*, Ed. I.I. Mueller, IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280, The Ohio State University, Columbus, U.S.A., pp. 93–97.
- SAVAGE, J.C. AND W.H. PRESCOTT (1973). Precision of geodetic distance measurements for determining fault movements. *J. Geophys. Res.* 78, pp. 6001–6008.
- SCHNEIDER, D. (1982). Complex crustal strain approximation. Department of Surveying Engineering Technical Report 91, University of New Brunswick, Fredericton, Canada.
- SNAY, R.A. AND J.G. GERGEN (1978). Monitoring regional crustal deformation with horizontal geodetic data. *Proc. 9th Geodesy/Solid Earth and Ocean Physics (GEOP) Research Conference, An International Symposium on the Applications of Geodesy to Geodynamics*, Ed. I.I. Mueller, IAG/IUGG and COSPAR, Columbus, U.S.A., October. Department of Geodetic Science Report 280, The Ohio State University, Columbus, U.S.A., pp. 87–92.

- TERADA, T. AND N. MIYABE (1929). Deformation of the earth crust in Kiransai District and its relation to the orographic feature. *Bulletin of the Earthquake Research Institute, University of Tokyo, Japan*, Vol. 7, pp. 223–241.
- TSUBOI, C. (1930). A note on the analytical treatments of the horizontal deformation of the earth crust. *Bulletin of the Earthquake Research Institute, University of Tokyo, Japan*, Vol. 8, pp. 384–392.
- TSUBOI, C. (1932). Investigation on the deformation of the earth's crust in the Tango District connected with the Tango earthquake of 1927 (Part 4). *Bulletin of the Earthquake Research Institute, University of Tokyo, Japan*, Vol. 10, pp. 411–436.
- TSUBOKAWA, I., T. DAMBARA AND A. OKADA (1968). Crustal movements before and after the Niigata earthquake. In: *General Report on the Niigata Earthquake of 1964*, Tokyo Electrical Engineering College Press.
- VACQUIER, V. AND R.E. WHITEMAN (1973). Measurement of fault displacement by optical parallax. *J. Geophys. Res.* 78 (5), pp. 858–865.
- VANIČEK, P. (1978). To the problem of noise reduction in sea level records used in vertical crustal movement detection. *Phys. of the Earth and Planetary Interiors* 17 (3), pp. 265–280.
- VANIČEK, P. (1980). Tidal corrections to geodetic quantities. NOAA Technical Report NOS 83 NGS 14, U.S. Department of Commerce, Rockville, U.S.A.
- VANIČEK, P. AND D. CHRISTODULIDES (1974). A method for the evaluation of vertical crustal movement from scattered geodetic relevelings. *Canad. J. Earth Sci.* 11 (5), pp. 605–610.
- VANIČEK, P., R.O. CASTLE AND E.I. BALAZS (1980). Geodetic leveling and its applications. *Rev. Geophys. and Space Phys.* 18 (2), pp. 505–524.
- VANIČEK, P., M.R. ELLIOTT AND R.O. CASTLE (1979). Four-dimensional modelling of recent vertical movements in the area of the southern California uplift. *Tectonophysics* 52, pp. 287–300.
- VANIČEK, P., K. THAPA AND D. SCHNEIDER (1981). The use of strain to identify incompatible observations and constraints in horizontal geodetic networks. *Manuscripta Geodaetica* VI(3), pp. 257–281.
- WHITTEN, C.A. (1967). Geodetic networks versus time. *Bull. Géod.* 84, pp. 109–116.
- ZSCHAU, J. (1976). Tidal sea load tilt of the coast and its application to the study of coastal and upper mantle structure. *Geophys. J. Roy. Astronom. Soc.* 44, pp. 577–593.

## AUTHOR INDEX

- Aardoom, L., 342, 447  
Abramowitz, M., 33, 52, 248, 284, 468, 476, 581  
Ackermann, F., 384, 447  
Adams, G.W., 340, 447  
Adler, R.K., 334, 453  
Ahlberg, J.H., 247, 284  
Airy, G.B., 134, 167  
Alberda, J.E., 413, 447  
Allan, A.L., 354, 443, 447  
Alt, J.N., 656  
Alterman, Z., 603, 656  
American Geophysical Union, 446, 447  
Anderle, R.J., 67, 167, 316, 323, 395, 447  
Anderson, E.G., 166, 167, 425, 427, 447  
Anderson, E.N., 419, 447  
Angus-Leppan, P.V., 405, 447  
Apparent Places of Fundamental Stars, 1979,  
    301, 447  
Arnold, K., 118, 167  
Ashkenazi, V., 409, 447  
Asimov, I., 3, 4, 52  
Associate Committee on Geodesy and Geo-  
    physics, 45, 52  
Avers, H.G., 78, 167
- Baarda, W., 215, 284, 411, 447  
Baeschlin, C.F., 368, 447  
Baker, T.F., 600, 656  
Balazs, E.I., 454, 656, 659  
Baldwin, A.L., 143, 168  
Balmino, G., 488, 496, 581, 583  
Bartelme, N., 412, 447  
Bayes, T., 215, 265, 284  
Beattie, D.S., 404, 447  
Beaumont, C., 604, 615, 656, 657  
Bender, P.L., 448, 634, 656  
Ben-Israel, A., 199, 284  
Berezin, I.S., 209, 284  
Beutler, G., 451  
Bibby, H.M., 655, 656  
Bjerhammar, A., 202, 284, 533, 581  
Bjerknes, V., 427, 448  
Blachut, T.J., 20, 52  
Blackman, R.B., 256, 284
- Blaha, G., 212, 274, 275, 284, 314, 381, 387, 448  
Blais, J.A.R., 388, 448  
Boal, J.D., 102, 169  
Bodemüller, H., 508, 581  
Böhm, J., 13, 52, 113, 167  
Bomford, G., 181, 284, 336, 350, 353, 396, 418,  
    432, 433, 448, 466, 496, 520, 581, 637, 656  
Bonnin, J., 169  
Boorstin, D.J., 15, 52  
Borre, K., 412, 413, 439, 448  
Bossler, J.D., 244, 265, 269, 274, 284, 344, 448  
Botting, D., 15, 52  
Bowditch, N., 417, 422, 448  
Bower, D.R., 605, 614, 656  
Bowie, W., 78, 167, 514, 582  
Bowring, B.R., 401, 454  
Bragard, L., 614, 656  
Brandenberger, A.J., 49, 52  
Bremmer, H., 421, 448  
Britting, K.R., 340, 448  
Brotén, N.W., 344, 448  
Brouwer, D., 311, 448  
Brown, D.C., 322, 387, 389, 448  
Brown, J.W., 30, 52  
Brown, L.A., 4, 8, 52  
Brown, L.D., 656  
Brown, R.D., 91, 167  
Brunavs, P., 421, 448  
Brunner, F.K., 432, 448, 646, 656  
Bruns, H., 375, 448, 498, 581  
Buchat, E., 115, 167  
Bullen, K.E., 72, 167  
Bunbury, E.H., 4, 8, 52  
Burford, R.O., 21, 53, 143, 171, 638, 641, 656  
Burg, J.P., 258, 284  
Burnside, C.D., 156, 167, 181, 284  
Bursa, M., 96, 107, 167, 394, 399, 448
- Cabaniss, G.H., 614, 656  
Cannon, J.B., 424, 448  
Capon, J., 258, 284  
Caputo, M., 555, 581  
Carrera, G., 329, 454  
Cassini, G., 78, 167

- Castle, R.O., 145, 172, 427, 448, 454, 625, 656, 659  
 Cazenave, A., 583  
 Celmins, A., 211, 284  
 Chamberlain, C., 405, 449  
 Champion, K.S.W., 170  
 Chandler, S.C., 66, 167  
 Chapman, S., 162, 163, 167  
 Chauvenet, W., 227, 284  
 Cheney, E.W., 194, 247, 248, 284  
 Cheney, R.E., 446, 448  
 Chi, S.C., 613, 656  
 Chinnery, M.A., 68, 172  
 Chovitz, B., 379, 448, 563, 581  
 Christodulides, D., 316, 448, 628, 629, 659  
 Chrzanowski, A., 52, 409, 448, 636, 656  
 Churchill, R.V., 30, 52  
 Cignoli, R., 195, 196, 284  
 Clark, D., 419, 448  
 Clark, J.A., 170  
 Clark, R.W., 16, 52  
 Clarke, A.R., 107, 167  
 Clemence, G.M., 311, 448  
 Coleman, R., 451, 563, 581  
 Comision Hidrologica de la Cuenca del Valle de México, 145, 167  
 Committee on Geodesy, 20, 47, 52, 442, 448  
 Condon, E.U., 157, 167, 594, 656  
 Conte, S.D., 364, 448  
 Cook, A.H., 22, 52, 181, 284, 584  
 Cooper, M.A.R., 181, 284  
 Coordinating Committee on Great Lakes Basic Hydraulic and Hydrologic Data, The, 614, 656  
 COSPAR, 153, 164, 167  
 Coulomb, J., 140, 167  
 Counselman, C.C., 343, 448  
 Craig, A.T., 25, 44, 53, 220, 226, 232, 234, 240, 285  
 Crawford, J.M., 286  
 Cross, P.A., 243, 285, 409, 447  
 Crow, E.L., 227, 285  
 Crowell, R.H., 54  
 Currie, R.G., 67, 167, 426, 449, 609, 657  
 Cyr, R.J., 167  
 Dambara, T., 659  
 Dare, P., 413, 449  
 Darling, F.W., 541, 583, 618, 657, 658  
 David, H., 227, 286  
 David, J.L., 658  
 Davies, P.C.W., 16, 52  
 Davis, F.A., 285  
 Davis, G.H., 146, 171  
 Davis, P.J., 194, 195, 246, 247, 285  
 de Boor, C., 364, 448  
 Delikaraoglou, D., 451, 454  
 Department of Energy, Mines and Resources, 79, 133, 168, 391, 399, 408, 415, 449, 536, 581  
 Department of Mines and Technical Surveys, 420, 449  
 Dermanis, A., 652, 657  
 de Sitter, W., 480, 581  
 de Vries, G., 647, 658  
 d'Hone, A., 268, 285  
 Diehl, W.S., 153, 168  
 Dijksterhuis, E.J., 4, 52  
 Dixon, W.J., 227, 285  
 Doodson, A.T., 129, 168, 589, 657  
 Dracup, J.F., 415, 449  
 Draper, C.S., 340, 449  
 Draper, N.R., 247, 285  
 Dreyer, J.L.E., 13, 52  
 Drude, P., 159, 168  
 Duerksen, J.A., 306, 450  
 Dufour, H.M., 412, 449  
 Dunn, P.J., 453  
 Durant, W., 8, 10, 11, 52  
 Eaton, R.M., 104, 169, 422, 449  
 Ebner, H., 384, 385, 449  
 Edge, R.C.A., 626, 657  
 Eggert, O., 353, 450  
 Eldred, R.J., 285  
 Elliott, M.R., 659  
 Emery, K.O., 425, 452  
 Encyclopaedia Britannica, 23, 52  
 Enochson, L., 249, 286  
 Environment Canada, 427, 449  
 Eremeev (Yeremeyev), V.F., 170, 327, 452, 455, 583  
 Esposito, P.B., 548, 581  
 Faddeev, D.K., 28, 52  
 Faddeeva, V.N., 28, 52  
 Fairbridge, R.W., 148, 168  
 Faller, J.E., 181, 285, 534, 581, 656  
 Farrell, W.E., 170, 603, 604, 657  
 Feller, W., 231, 285  
 Fila, K., 405, 449  
 Fischer, I., 566, 574, 581  
 Fite, E.D., 11, 52  
 Flanders, H., 34, 52  
 Forsberg, R., 545, 574, 584  
 Forward, R.L., 181, 285, 508, 581  
 Francheteau, J., 169  
 Frank, F.C., 647, 655, 657  
 Freeman, A., 11, 52

- Freund, J.E., 44, 52  
 Frost, N.H., 146, 168  
 Fubara, D.M.J., 379, 449
- Gale, L.A., 146, 168  
 Gaposhkin, E.M., 80, 113, 114, 119, 168, 343, 451, 559, 560, 563, 581  
 Garfinkel, B., 305, 449  
 Garland, G.D., 82, 168  
 Gass, I.G., 97, 130, 140, 168  
 Gauss, K.F., 214, 285  
 Geldart, L.P., 54, 584  
 Gergen, J.G., 648, 658  
 Gibson, J.R., 453  
 Gilbert, G.K., 137, 168  
 Girnius, A.G., 447  
 Goad, C.C., 344, 448, 449, 605, 657  
 Godin, G., 125, 127, 168, 249, 256, 285  
 Gold, B., 249, 285  
 Goldman, S., 184, 249, 251, 285  
 Goldreich, P., 119, 168  
 Goodkind, J.M., 534, 582, 609, 615, 657  
 Gore, R., 578, 584  
 Gough, D.I., 132, 168  
 Gough, W.I., 132, 168  
 Gourevitch, S.A., 343, 448  
 Graber, M.A., 67, 168  
 Grace, H., 144, 172  
 Grafarend, E.W., 243, 244, 268, 284, 285, 434, 454  
 Grant, S., 422, 454  
 Graybill, F.A., 219, 240, 285  
 Greenberg, M.D., 35, 37, 52  
 Greenwood, J.B., 181, 285  
 Gregerson, L.F., 335, 340, 449, 534, 574, 582  
 Greville, T.N.E., 199, 284  
 Grotens, E., 545, 582  
 Groueff, S., 8, 13, 52  
 Guier, W.H., 181, 285  
 Guinot, B., 68, 168, 169  
 Guyenne, T.D., 20, 53
- Hadley, G., 201, 205, 285  
 Hagihara, Y., 24, 52  
 Halmos, F., 413, 449  
 Hamilton, A.C., 20, 53, 146, 171, 439, 454  
 Hamilton, W.C., 219, 220, 227, 232, 235, 237, 239, 285  
 Hancock, H., 31, 53, 205, 285  
 Hanson, R.H., 209, 285  
 Hanson, R.J., 199, 208, 285  
 Hapgood, C.H., 10, 53  
 Hardy, R.L., 629, 657  
 Harkrider, D.G., 604, 657
- Harland, P., 243, 285  
 Harman, H.H., 413, 450  
 Haurwitz, B., 163, 168  
 Hayford, J.F., 113, 114, 143, 168, 514, 582  
 Heiskanen, W.A., 96, 106, 107, 114, 115, 130, 137, 168, 326, 351, 365, 366, 450, 466, 500, 503, 512, 514, 520, 524, 526, 532, 582  
 Hela, I., 105, 168, 427, 450  
 Helmert, F.R., 45, 53, 369, 402, 450, 566, 574, 582  
 Hendershott, M.C., 129, 168  
 Henriksen, S.W., 119, 168  
 Hieber, S., 20, 53  
 Hill, M.N., 105, 129, 156, 161, 168, 442, 450  
 Hirvonen, R.A., 117, 168, 202, 285, 533, 538, 582  
 Hobson, E.W., 468, 469, 472, 473, 476, 519, 582  
 Hochstadt, H., 470, 582  
 Hodges, D.J., 181, 285  
 Hogg, R.V., 25, 44, 53, 220, 226, 232, 234, 240, 285  
 Hoheisel, G., 37, 53  
 Holdahl, S.R., 145, 146, 168, 432, 450, 627, 629, 657  
 Holland, G.L., 609, 657  
 Hollwey, J.R., 447  
 Hopfield, H.A., 315, 450  
 Hoskinson, A.J., 306, 450  
 Hothem, L.D., 395, 450  
 Hotine, M., 156, 168, 334, 358, 359, 375, 380, 395, 450  
 Hradilek, L., 380, 450  
 Huggett, G.R., 336, 450  
 Hunziker, E., 92, 169, 508, 582  
 Hursh, J.W., 444, 450  
 Hydrographer of the Navy, 104, 169, 441, 450
- Ingham, A.E., 419, 441, 450  
 International Association of Geodesy, 73, 74, 79, 88, 113, 114, 165, 169, 480, 483, 535, 582  
 International Astronomical Union, 71, 169, 589, 657  
 International Union of Geodesy and Geophysics, 47, 53, 476, 485, 582  
 Irving, E., 139, 169  
 Isner, J.R., 405, 407, 450  
 Iyer, H.M., 584
- Jacchia, L.G., 22, 53  
 Jachens, R.C., 615, 617, 657  
 Jaeger, J.C., 649, 653, 657  
 Jarosch, H., 656  
 Jaswon, M.A., 37, 53  
 Jeffreys, H., 66, 67, 119, 169, 215, 285  
 Jekeli, C., 543, 582

- Johler, J.R., 421, 450  
 Jones, H.E., 399, 450  
 Jordan, W., 353, 450  
 Jorgensen, P.S., 315, 450  
 Jurkowski, G.A., 656
- Kääriäinen, E., 626, 657  
 Kádár, I., 412, 449  
 Kalman, R.E., 277, 285  
 Kasahara, K., 647, 648, 657  
 Katinas, G., 452  
 Kaula, W., 62, 91, 104, 119, 169, 311, 389, 451, 538, 545, 550, 554, 555, 556, 558, 582  
 Kayton, M., 338, 451  
 Kellar, W.G., 450  
 Keller, M., 383, 451  
 Kelm, R., 284  
 Keys, D.A., 54, 584  
 King, R.W., 169  
 King-Hele, D.G., 52, 549, 582  
 Kirkham, B.P., 451  
 Klosko, S.M., 583  
 Knight, W., 209, 283, 285, 403, 407, 451  
 Kobold, F., 92, 169, 508, 582  
 Koch, K.R., 486, 496, 582  
 Kochin, N.E., 36, 53  
 Kolaczek, B., 380, 451  
 Kolenkiewicz, R., 453  
 Konecny, G., 409, 448  
 Korhonen, J., 623, 657  
 Korn, G.A., 25, 33, 53, 205, 285, 326, 451  
 Korn, T.M., 25, 33, 53, 205, 285, 326, 451  
 Kouba, J., 102, 169, 323, 343, 373, 451, 455, 462, 582  
 Kovalevsky, J., 58, 169, 547, 550, 582  
 Krakiwsky, E.J., 20, 46, 53, 54, 282, 285, 286, 322, 323, 373, 395, 430, 451, 453, 455  
 Krarup, T., 261, 285, 533, 582  
 Krebs, O.A., 148, 168  
 Kreyszig, E., 31, 53  
 Krogstad, R.S., 287  
 Kukkamäki, T.J., 137, 169, 373, 430, 431, 451, 454  
 Kumar, M., 394, 452
- Lachapelle, G., 539, 545, 575, 578, 580, 582, 583, 584  
 Lambeck, K., 399, 451, 549, 559, 560, 581, 583, 607, 657  
 Lambert, A., 604, 615, 656, 657, 658  
 Lambert, W.D., 503, 541, 583, 618, 653  
 Lanczos, C., 252, 285  
 Landkof, N.S., 36, 53, 85, 169  
 Langley, R.B., 69, 169, 344, 451, 454
- Latimer, J.H., 343, 451  
 Laubscher, R.E., 583  
 Lawson, C.L., 199, 208, 285  
 Ledersteger, K., 118, 169  
 Lee, L.P., 363, 451  
 Lee, W.H.K., 104, 169  
 Legg, T.H., 448  
 Lehr, C.G., 181, 285  
 Leitao, C.D., 452  
 Le Méhauté, B., 160, 169  
 Lennon, G.W., 22, 53, 181, 286, 600, 604, 656, 658  
 LePichon, X., 139, 169  
 Lerch, F.J., 580, 583  
 Levallois, J.J., 369, 451, 483, 583  
 Levine, J., 637, 656, 658  
 Liebelt, P.B., 184, 189, 264, 286  
 Lilly, J.E., 146, 168  
 Lindzen, R.S., 162, 163, 167  
 Lisitzin, E., 105, 168, 427, 450  
 List, R.J., 153, 154, 169  
 Livieratos, E., 652, 657  
 Locke, J.L., 448  
 Longman, I.M., 604, 658  
 Love, A.E.H., 590, 593, 594, 658  
 Lucas, J., 581  
 Lucht, H., 434, 435, 451  
 Luenberger, D.G., 196, 286
- MacDonald, G.F.K., 66, 68, 69, 170, 600, 658  
 MacDoran, P.F., 345, 451  
 MacMillan, D.H., 129, 169  
 MacMillan, W.D., 36, 53, 63, 72, 74, 83, 85, 165, 169, 464, 474, 583  
 MacPhee, S.B., 422, 442, 451  
 Magness, T.A., 237, 286  
 Maling, D.H., 334, 359, 451  
 Malone, T.F., 162, 169  
 Mamon, G., 450  
 Mansinha, L., 66, 169  
 Markowitz, W., 68, 69, 169  
 Marsh, J.G., 171, 446, 448, 584  
 Marsh, P.W., 656  
 Masry, S.E., 453  
 Mather, R.S., 167, 327, 445, 451, 520, 563, 581, 583  
 Matthews, D.J., 442, 452  
 Mavko, G., 625, 658  
 Maxfield, M.W., 285  
 Maynes, J.H.B., 447  
 McConnell, A.J., 41, 53  
 McCulloh, T.H., 500, 583  
 McGoogan, J.T., 446, 452  
 McGuire, J.B., 237, 286

- McKeown, D.L., 104, 169  
 McLellan, C.D., 403, 452  
 McWhirter, N., 128, 132, 170  
 McWhirter, R., 128, 132, 170  
 Meade, B.K., 636, 637, 658  
 Meade, R.H., 425, 452  
 Medallion World Atlas, 154, 170  
 Meissl, P., 407, 412, 414, 439, 447, 448, 452  
 Melchior, P., 62, 65, 118, 119, 128, 130, 170, 181,  
     286, 590, 591, 592, 595, 658  
 Menard, H.W., 143, 170  
 Mendes, G.M., 285  
 Menzel, D.H., 154, 156, 157, 162, 170, 318, 452  
 Mepham, M.P., 283, 285, 403, 407, 451  
 Merry, C.L., 89, 96, 117, 170, 171, 352, 428, 452,  
     453, 454, 543, 568, 570, 574, 577, 583, 584  
 Messih, F.Z.A., 439, 454  
 Mikhail, E.M., 202, 208, 244, 270, 286  
 Miller, A.R., 425, 452  
 Miller, L.S., 452  
 Miller, R.G., 231, 286  
 Miller, R.W., 642, 658  
 Minzner, R.A., 153, 170  
 Misner, C.W., 147, 170  
 Mitra, S.K., 199, 275, 286  
 Miyabe, N., 651, 659  
 Moffett, J.B., 316, 452  
 Molodenskij, M.S., 117, 170, 372, 394, 452, 486,  
     495, 498, 528, 531, 532, 543, 562, 573, 583  
 Montgomery, R.B., 425, 452  
 Moore, C.H., 554, 584  
 Morando, B., 549, 583  
 Morelli, C., 534, 583  
 Moritz, H., 65, 106, 114, 115, 119, 168, 170, 184,  
     277, 286, 326, 351, 365, 366, 450, 466, 500,  
     503, 520, 524, 532, 533, 537, 545, 582, 583  
 Morrey, C.B., Jr., 31, 39, 53  
 Morrison, F., 486, 496, 581, 582, 583  
 Morrison, N., 29, 53, 277, 281, 286  
 Morrison, N.L., 146, 168, 627, 657  
 Moss, R.W., 287  
 Mueller, I.I., 20, 53, 61, 62, 132, 170, 181, 286,  
     299, 300, 303, 305, 306, 307, 308, 309, 316,  
     335, 373, 389, 394, 395, 430, 451, 452, 500,  
     508, 583  
 Müller, K., 435, 452  
 Munk, W.M., 66, 68, 69, 170, 600, 658  
 Murty, T.S., 609, 657  
 Myint-U, T., 519, 583  
 Nagy, D., 80, 170  
 Nakagawa, J., 147, 170  
 Nason, R.D., 656  
 Nassar, M.M., 430, 452, 453  
 Nassau, J.J., 58, 170  
 National Aeronautics and Space Administration,  
     21, 22, 53, 99, 100, 170, 344, 452, 634, 658  
 Newcomb, S., 62, 66, 170, 299, 452  
 Newton, I., 70, 170  
 Ng, A.T.Y., 548, 581  
 Nickerson, B.G., 245, 286, 451  
 Niell, A.E., 451  
 Nilson, E.N., 284  
 IXth National Surveying Teachers' Conference,  
     50, 53  
 Niskanen, E., 526, 582  
 Nordenskjöld, A.E., 11, 53  
 Norrie, D.H., 647, 658  
 Nur, A., 625, 658  
 Nyland, E., 655, 658  
 Odishaw, H., 157, 167, 594, 656  
 Officer, C.B., 130, 170, 486, 583  
 Okada, A., 659  
 O'Keefe, J.A., 119, 169  
 Ong, K.M., 451  
 Orlin, H., 539, 583  
 Orlov, A. Ya., 68, 170  
 Otnes, R.K., 249, 286  
 Ozawa, I., 591, 658  
 Pannekoek, A., 8, 53  
 Parthasarathy, K.R., 184, 286  
 Paul, M.K., 327, 452  
 Pedersen, G.P.M., 66, 170  
 Pekeris, C.L., 656  
 Pelikán, M., 158, 170  
 Peltier, W.R., 132, 136, 170  
 Perelmuter, A., 274, 286  
 Permanent Service for Mean Sea Level, 105, 170  
 Peterson, A.E., 452  
 Petterssen, S., 151, 152, 163, 170  
 Picha, J., 170, 452, 583  
 Pick, M., 82, 170, 327, 452, 486, 583  
 Poland, J.F., 146, 171  
 Pollard, H., 554, 584  
 Pond, H.L., 170  
 Pope, A.J., 207, 211, 227, 228, 229, 237, 238, 286,  
     646, 647, 658  
 Pratt, J.H., 133, 171  
 Prescott, W.H., 636, 658  
 Prey, A., 104, 171  
 Price, J.J., 34, 52  
 Protter, M.H., 31, 39, 53  
 Quesenberry, C., 227, 286  
 Quraishee, G.S., 150, 171

- Rader, C.M., 249, 285  
 Rainsford, H.F., 353, 452  
 Ramsayer, K., 335, 452  
 Ramsden, S.A., 52  
 Rao, C.R., 199, 275, 286  
 Rapp, R.H., 89, 171, 445, 453, 537, 538, 539, 561,  
     562, 584  
 Rappleye, H.S., 366, 432, 452  
 Reid, D.B., 442, 453  
 Reilinger, R.E., 656  
 Rektorys, K., 33, 53  
 Remmer, O., 434, 439, 453  
 Remondi, B.W., 344, 449  
 Resch, G.M., 451  
 Revuz, D., 185, 286  
 Richardus, P., 334, 453  
 Rikitake, T., 141, 171, 625, 653, 658  
 Rinne, K., 20, 45, 53  
 Rizos, C., 167, 451  
 Robbins, A.R., 52, 181, 286, 296, 298, 307, 353,  
     453  
 Robertson, D.S., 450  
 Rochester, M.G., 66, 170, 171  
 Roden, G.I., 425, 453  
 Roesler, G., 167  
 Rosen, R.D., 169  
 Rossby, C.G., 163, 171  
 Rossiter, J.R., 426, 453  
 Routh, E.J., 24, 53  
 Rummel, R., 445, 453  
 Runcorn, S.K., 130, 171  
 Saastamoinen, J.J., 52, 305, 316, 336, 453  
 St. George, E., Jr., 167  
 Salstein, D.A., 169  
 Sandström, J.W., 427, 448  
 Santerre, R., 451  
 Savage, I.R., 215, 286  
 Savage, J.C., 21, 53, 143, 171, 635, 636, 656, 658  
 Scarborough, J.B., 339, 453  
 Schild, A., 40, 54, 196, 187  
 Schmid, E., 265, 281, 286  
 Schmid, H.H., 102, 171, 265, 281, 286, 395, 453  
 Schmidt, K., 184, 286  
 Schneider, D., 454, 649, 658, 659  
 Schneider, E., 435, 452  
 Schut, G.H., 383, 453  
 Schwarz, C.R., 270, 286, 387, 404, 453  
 Schwarz, K.P., 110, 113, 171, 539, 584  
 Seppelin, T.O., 113, 114, 171, 563, 584  
 Shapiro, I.I., 169  
 Sheriff, R.E., 54, 584  
 Shirley, R., 581  
 Siegel, S., 215, 286  
 Simmons, L.G., 99, 171  
 Simonsen, O., 373, 453  
 Singer, I., 195, 196, 286  
 Slater, L.E., 336, 450  
 Slowey, J.W., 22, 53  
 Slutsky, M., 581  
 Small, J.B., 636, 658  
 Smart, W.M., 62, 171, 301, 453  
 Smith, D.E., 314, 316, 448, 453  
 Smith, H., 247, 285  
 Smith, J.R., 181, 286  
 Smith, P.J., 168  
 Smylie, D.E., 66, 169  
 Snay, R.A., 407, 453, 648, 658  
 Sodano, E.M., 353, 453  
 Sollins, A., 541, 584  
 Soltz, J.A., 450  
 Stacey, F.D., 584  
 Stanley, H.R., 446, 453  
 Stearn, J.L., 646, 647, 658  
 Steeves, P., 209, 285  
 Steeves, R.R., 244, 286  
 Stefansky, W., 229, 286  
 Stegun, I.A., 33, 52, 248, 284, 468, 476, 581  
 Stettner, H.S., 658  
 Stoch, L., 308, 453  
 Stokes, G.G., 516, 584  
 Stommel, H., 148, 171  
 Strange, W.E., 167, 171, 450, 500, 584  
 Stuifbergen, N., 449  
 Sugimura, 648, 657  
 Symm, G.T., 37, 53  
 Symon, K.R., 63, 171, 549, 584  
 Syngle, J.L., 40, 54, 196, 287  
 Syverson-Krakiwsky, M.L., 286  
 Takeuchi, H., 170  
 Telford, W.M., 21, 54, 534, 584  
 Terada, T., 651, 659  
 Terrien, J., 156, 171  
 Thapa, K., 243, 285, 454, 659  
 Theil, H., 267, 287  
 Thomas, P.D., 353, 362, 453  
 Thompson, E.H., 28, 29, 38, 54, 209, 273, 275,  
     287  
 Thompson, M.M., 443, 453  
 Thompson, W., 229, 287  
 Thomson, D.B., 286, 328, 395, 413, 420, 442, 451,  
     453, 455  
 Thomson, D.W., 8, 54  
 Thorne, K.S., 170  
 Thorson, C.W., 307, 454  
 Tienstra, J.M., 282, 287  
 Tobey, W.M., 402, 454

- Tompkins, P., 3, 54  
Toomre, A., 119, 168  
Torrence, M.H., 453  
Tropper, A.M., 37, 53  
Trotter, H.F., 54  
Trotter, J.E., 389, 448  
Tscherning, C.C., 538, 545, 574, 584  
Tsiang, C.R.H., 285  
Tsuboi, C., 141, 171, 635, 652, 653, 659  
Tsubokawa, I., 625, 659  
Tucker, R.H., 553, 584  
Tukey, J.W., 256, 284
- U.S. Army Topographic Command, 98, 100, 171  
U.S. Department of Commerce, 399, 454  
U.S. Federal Geodetic Control Committee, 434, 454
- Vacquier, V., 637, 659  
Vali, V., 181, 287  
Valliant, H.D., 534, 584  
Van den Hout, C.M.A., 384, 454  
Vanicek, P., 20, 21, 46, 53, 54, 67, 68, 89, 96, 117, 146, 150, 170, 171, 256, 287, 321, 329, 330, 344, 352, 377, 413, 414, 425, 426, 427, 430, 434, 439, 448, 449, 451, 452, 454, 543, 568, 570, 574, 577, 583, 584, 599, 604, 609, 612, 615, 616, 618, 628, 629, 632, 652, 658, 659  
Veis, G., 181, 287, 389, 394, 447, 454, 554, 584  
Vening Meinesz, F.A., 96, 130, 135, 137, 168, 171, 512, 514, 522, 526, 582, 584  
Vignal, J., 373, 430, 454  
Vincent, S., 89, 167, 171, 523, 580, 584  
Vincenty, T., 375, 376, 378, 401, 454  
von Arx, W.S., 419, 422, 454  
Vyskocil, V., 170, 452, 583
- Wagner, C.A., 583  
Walcott, R.J., 131, 137, 171  
Walsh, J.L., 284  
Walters, L.C., 450  
Waslef, A.M., 439, 454  
Watts, D.G., 256, 257, 258, 285  
Wegener, A., 138, 171  
Weiffenbach, G.C., 160, 172, 181, 285, 380, 451
- Wells, D.E., 156, 172, 315, 316, 322, 330, 343, 397, 420, 421, 422, 442, 448, 449, 451, 453, 454, 455  
Wells, F.J., 68, 172  
Wells, H.G., 5, 11, 54  
Wells, W.T., 452  
Westerfield, E.E., 342, 455  
Wheeler, J.A., 170  
Whiteman, R.E., 637, 659  
Whitten, C.A., 141, 172, 648, 659  
Wilks, S.S., 215, 258, 287  
Will, L.S., 147, 172  
Williamson, R.E., 31, 54  
Willke, T.A., 227, 287  
Wilson, G., 144, 172  
Wilson, R.C.L., 168  
Wise, P.J., 444, 455  
Wolf, H., 405, 455  
Wolf, P., 382, 455  
Wong, L., 578, 584  
Wonnacott, R.J., 42, 43, 54, 215, 222, 227, 268, 287  
Wonnacott, T.H., 42, 43, 54, 215, 222, 227, 268, 287  
World Almanac and Book of Facts 1984, 152, 172  
World of Learning 1981-82, 47, 49, 54  
Worsley, G., 342, 455  
Wrede, R.C., 31, 54, 196, 287  
Wyatt, P., 581  
Wylie, C.R., Jr., 37, 54
- Yeremeyev (Eremeev), V.F., 170, 327, 452, 455, 583  
Yerkes, R.F., 145, 172  
Yionoulis, S.M., 315, 455  
Young, G.M., 405, 450  
Yumi, S., 67, 172  
Yurkina, M.I., 170, 327, 452, 455, 583
- Zakatov, P.S., 349, 352, 455  
Zhidkov, N.P., 209, 284  
Zschau, J., 604, 659  
Zygmund, A., 252, 287

## SUBJECT INDEX

- a posteriori variance factor 213  
aberration  
    annual 301  
    diurnal 301  
absolute confidence ellipse 355  
absolute confidence ellipse for displacement 641  
absolute confidence ellipsoid 347  
absolute constraint 270  
absolute deflection of vertical 91  
absolute extremum 31  
absolute geoidal height 89  
absolute gravity 593  
absolute gravity measurement 534  
absolute value of complex number 30  
absolute vertical displacement 619  
abstract elasticity of network 413  
accelerated (crustal) movement 625  
accelerated variations of sea level 632  
acceleration  
    Coriolis 162  
    pressure gradient 162  
    tidal 125  
accelerometer 337  
accuracy 42  
    relative 42  
acoustic positioning system 422  
actual polar motion 66  
adjusted parameter 205  
adjustment 258  
    block 383  
    generalized 265  
    simple 258  
stereomodel 384  
    semi-analytical 384  
stereomodel block 384  
    two-component 259  
adjustment of horizontal network 400  
adjustment of leveling network 429  
adjustment of three-dimensional network 375  
air density 153  
air drag 548  
air humidity 155  
air pressure 153  
air temperature 152  
airborne profile recording 443  
almucantar 293  
alternative hypothesis 220  
altimeter footprint 444  
altimetry: satellite 444  
amplitude of complex number 30  
amplitude of trigonometrical term 253  
analysis  
    data series 189  
    harmonic 252  
    multivariate 232  
    multivariate subset 232  
    singular value 208  
    spectral 252  
    vector 31  
analysis of trend 245  
analytical mechanics 549  
analytical model of distortion 415  
anchored object 416  
angle  
    datum misalignment 327  
    ellipsoidal horizontal 355  
    geodetic vertical 328  
    horizontal 181  
        plane 356  
        projected 356  
    hour 302  
    refractive 158  
    to star  
        horizontal 181  
        vertical 181  
    vertical 294  
angular-momentum integral 547  
angular velocity 65  
annual aberration 301  
annual parallax 300  
annual variation of sea level 149  
anomalous uplift 146  
anomalous subsidence 146  
anomaly 310  
    eccentric (orbital) 310  
gravity *see* gravity anomaly

- height 118
- mean (orbital) 310
- true (orbital) 310
- anti-root of oceanic block 135
- antisymmetrical deformation 654
- antisymmetrical matrix 26
- AP *see* apparent place coordinate system
- APFS *see* Apparent Places of Fundamental Stars
- aphelion 58
- apogee 310
- apparent place coordinate system 298
- Apparent Places of Fundamental Stars 301
- approximant 246
- approximation
  - mean quadratic 247
  - problem of 247
  - spline 247
  - uniform 247
- APR *see* airborne profile recording
- arc
  - orbital 319
  - short 387
  - long 388
- arc length 39
- arc-to-chord ( $T-t$ ) correction 362
- argument: vector 26
- argument of complex number 30
- argument of perigee 311
- ascending node 311
  - right ascension of 311
- assessment of systematic network distortion 439
- asthenosphere 130
- astrodeflection (astro-geodetic deflections or astronomical deflections) 564
- densification of observed 573
- astro-geodetic deflections 564
- astro-geodetic geoid 566
- astro-geodetic levelling 566
- astro-gravimetric geoid 577
- astronomical azimuth 294
  - instantaneous 298
- astronomical coordinate system:
  - local (LA) 294
- astronomical deflection of vertical 564
- astronomical determination of marine position 422
- astronomical latitude 296
- astronomical longitude 296
- astronomical meridian 293
- astronomical positioning (§15.2)
- astronomical radio-interferometry 344
- astronomical refraction 302
- astronomical zenith distance 294
- astronomy 22
- geodetic (§15.1)
- atmosphere 151
- atmospheric correction to gravity
  - first order 164
  - second order 166
- atmospheric correction to gravity potential 164, 166
- atmospheric dynamics 161
- atmospheric pressure effect on sea level 425
- atmospheric science 22
- atmospheric tide 163
- atomic time 303
- attraction: gravitational 71
- attraction of tidal water 600
- autocovariance function 186
- auxiliary model 183
- average differential rotation 655
- axiom of metric 194
- axis of strain 653
- azimuth
  - astronomical 294
  - instantaneous 298
  - ellipsoidal 350
  - geodetic 328
  - grid 362
  - inverse 354
  - Laplace 348
  - projected 361
- azimuth determination by hour angles of stars near culmination 308
- azimuth determination by hour angles of stars near elongation 309
- azimuth equation Laplace 331
- azimuth of chord 362
- back distribution of residuals 437
- balance: torsion 499
- band-pass filter 257
- barometric equation: Laplace 443
- barometric height 443
- barometric pressure loading 133
- base
  - orthogonal 247
  - orthonormal 248
- base function 246
- base line: geodetic 637
- basic postulate of mathematical statistics 217
- bathymetry 441
- Bayes filter 281
- Bayesian approach to weighting 268
- Bayesian statistics 215
- bench mark 98
  - reference 424
- best solution 194

- bias 42  
**biaxial ellipsoid** 110  
**BIH** *see* Bureau International de l'Heure  
**binormal vector** 39  
**bivariate covariance matrix** 232  
**bivariate probability density function** 232  
**bivariate population mean** 232  
**block**  
  continental 135  
  innermost 541  
  oceanic 135  
  outer 543  
  photogrammetrical 381  
**block adjustment**  
  bundle 383  
  stereomodel 384  
**blocking**: Helmert 405  
**body of the earth**: normal 87  
**body tide** 590  
**Bonferroni inequality** 231  
**Bouger gradient** 502  
**Bouguer gravity anomaly** 513  
**Bouguer plate gradient** 502  
**boundary** *see* tectonic plate boundary  
**boundary demarcation** 20  
**boundary value** 35  
**boundary value problem** 35  
  Dirichlet 36  
  geodetic 517  
    Stokes's solution to (§22.1)  
    Molodenskij's solution to (§22.2)  
  Neumann 36  
  Sturm-Liouville 35  
**Brillouin sphere** 467  
**broadcast ephemeris** 315  
**Brun's formula** 493  
  generalized 493  
**bundle block adjustment** 383  
**Bureau International de l'Heure** 66
- Canonical equation of motion** 550  
**Cartesian coordinate system** 37  
**Cauchy-Riemann equation** 358  
**caving in** 146  
**celestial coordinate system** 293  
**celestial equator** 293  
**celestial horizon** 293  
**celestial mechanics** 547  
**celestial parallel** 293  
**celestial pole**  
  north 292  
  south 292  
**celestial sphere** 292  
**central quadric** 29
- centrifugal force 73  
  centrifugal potential 83  
  chain: Markov 185  
  Chandler period 66  
  change of gravitational constant 147  
  chart  
    corange 129  
    cotidal 129  
    hydrographic 441  
  chi-squared goodness of fit test 227  
  chi-squared probability density function 42  
  chi-squared test on variance 227  
  Cholesky: method of 209  
  chord: azimuth of 362  
  chord length 356  
  CIO *see* conventional international origin  
  circle: vertical 293  
  circular orbit 550  
  circular point 41  
  Clairaut equation 118  
  Clairaut theorem 114  
  clock offset 314  
  close satellite 546  
  closed curve 34  
  closed loop 368  
  closed surface 34  
  closest approach: point of (PCA) 318  
**coefficient**  
  correlation 44  
  normal potential 480  
  potential 471  
  truncation 543  
**coefficient matrix** 205  
**cogenoid** 525  
**collocation**: least squares 261  
**combination of models** 181  
**combined design problem** 243  
**combined positioning and potential coefficient determination** 563  
**compact neighbourhood of point** 30  
**compact space**: locally 30  
**compaction**: ground 143  
**comparison of horizontal positions** 638  
**complete solution** 35  
**complex Fourier transformation** 255  
**complex function** 30  
**complex number** 29  
  absolute value of 30  
  amplitude of 30  
  argument of 30  
  conjugate of 30  
  imaginary part of 29  
  real part of 29  
**component**

- datum translation 327
- deflection of vertical 92
- systematic 183
- composite statistical hypothesis 220
- computer science 23
- commutative diagram 38
- condition
  - minimum distance 194
  - minimum norm 194
- condition model 179
  - linear 179
- conditional probability 44
- conditions for parallelism 331
- confidence ellipse
  - absolute 355
  - point 355
    - standard 355
  - relative 355
  - simultaneous 408
- confidence ellipse for displacement
  - absolute 641
  - relative 642
- confidence ellipsoid 346
  - absolute 347
  - point 346
    - standard 347
  - relative 347
  - simultaneous 392
- confidence interval ( $1-\alpha$ ) 224
  - point
    - out-of-context 439
    - (simultaneous) in-context 439
  - relative 439
- confidence level 221
  - critical 223
- confidence region 222
- configuration of satellite positions 314
- conformal mapping 357
- conformal mapping plane 401
- conformality 357
- congruity 26
- conjugate of complex number 20
- conservative field 82
- constant 177
  - Doodson tidal 589
  - gravitational 71
  - integration 35
  - precessional 299
- constant acceleration displacement model 647
- constant vector 178
- constrained displacement model
  - spatially 647
  - temporally 647
- constrained magnitude displacement model 647
- constraint 269
  - absolute 270
  - inner 274
  - minimal 273
  - weighted 270
- constraint function 180
- continental block 135
- continuous representation of earth surface 104
- contraction 652
- control surveying 19
- conventional international origin 66
- conventional terrestrial coordinate system (CT) 296
- convergence: meridian 360
- converging tectonic plate boundary 141
- convolution 33
- convolution integral 33
- convulsive filter 249
- coordinate 37
- coordinate system
  - apparent place (AP) 298
  - Cartesian 37
  - celestial *see* celestial coordinate system
  - conventional terrestrial (CT) 296
  - curvilinear 38
  - ecliptical 302
  - ellipsoidal (EL)
    - geodetic 38
    - one-parametric (EL) 464
  - family of 38
  - generalized 549
  - geocentric 296
  - geodetic (G) 324
    - local (LG) 328
  - heliocentric 293
  - inertial 302
  - instantaneous terrestrial (IT) 297
  - local astronomical (LA) 294
  - locally orthonormal 39
  - mapping (M) 334
  - non-parametric 38
  - orbital (OR) 311
  - orientation of 37
  - origin of 37
  - polar 450
  - polarity of 37
  - right ascension (RA)
    - mean (MRA) 294
    - true (TRA) 300
  - spherical 38
  - surface 40
  - terrestrial *see* terrestrial coordinate system
  - topocentric 295
  - transformation between Cartesian 37

- corange chart 129  
 Coriolis acceleration 162  
 correction  
     annual parallax 300  
     arc-to-chord ( $T-t$ ) 362  
     complete azimuth 350  
     (propagation) delay *see* refraction correction  
     distance 352  
     dynamic (height) 370  
     free air 77  
     height of target 350  
     horizontal angle 351  
     horizontal direction 350  
     ionospheric delay 315  
     ionospheric Doppler count 322  
     Laplace 348  
     levelling refraction 431  
     loading 606  
     normal (height) 429  
     normal section to geodesic 350  
     orthometric (height) 429  
     refraction *see* refraction correction  
     skew-normal 349  
      $T-t$  (arc-to-chord) 362  
     tidal *see* tidal variation  
     tropospheric delay 315  
     tropospheric Doppler count 321  
     correction vector (in adjustment) 205  
 correlate: Lagrange 205  
 correlation 44  
 correlation coefficient 44  
 coseismic displacement 141  
 cotidal chart 129  
 count  
     (integrated) Doppler 320  
     relative 42  
 course: ship 416  
 covariance 43  
 covariance filter 261  
 covariance function 186  
 covariance law 197  
 covariance matrix 43  
     bivariate 232  
 critical confidence level 223  
 critical configuration of satellite positions 314  
 critical significance level 223  
 cross-covariance function 189  
 cross-covariance matrix 211  
 crossover point 445  
 crust: earth's 130  
 crustal loading §8.2  
 crustal movement  
     accelerated 625  
     horizontal Ch. 8  
     vertical Ch. 8  
         linear 622  
 crustal velocity surface 628  
     standard deviation of 629  
 CT *see* conventional terrestrial coordinate system  
 culmination 306  
     star near 308  
 currents: sea 425  
 curvature 39  
     of the plumbline §21.3  
     radius of 40  
         meridian 111  
         prime vertical 324  
 curvature gradient 502  
 curve  
     closed 34  
     geodesic 40  
     pedal 357  
     spatial 39  
 curvilinear coordinate system 38  
 data series 245  
     smoothed 249  
 data series analysis 189  
 datum 40  
     chart 441  
     geodetic *see* horizontal and/or vertical (geodetic) datum  
     horizontal geodetic *see* horizontal (geodetic) datum  
     sounding 441  
     vertical geodetic 98  
 datum (position) parameter 327  
     geocentric 327  
     topocentric 330  
 datum positioning by floating datum technique 399  
 datum positioning by set of selected points 400  
 datum positioning by standard technique 396  
 datum misalignment angle 327  
 datum translation component 327  
 day  
     sidereal 59  
     solar 59  
 declination 293  
 decomposition: orthogonal 199  
 decomposition formula: Legendre 473  
 decomposition of observable 187  
 decomposition of random series 264  
 defect of matrix 28  
 deflection of vertical 91  
     absolute 91  
     astro-geodetic (astronomical) 564  
     components of 92

- geoidal 491
  - refined 575
- gravimetric 573
  - incomplete 573
- meridian component of 93
- Molodenskij 491
  - prediction of 573
- prime vertical component of 93
- relative 91
  - surface 91, 491
- tidal variation of astronomical 597
  - topographical noise in 568
- deflection point 565
- deformation
  - antisymmetrical 654
  - crustal loading 130, §8.2
  - landslide 146
  - man made 143, §8.4
  - polar motion 147
  - symmetrical 649
  - tectonic 138, §8.3
  - tidal 124
    - varying earth's spin velocity induced 609
- deformation potential 592
- degree of freedom 213
- Delauney coordinates 550
- delay correction
  - ionospheric 315
  - tropospheric 315
- demarcation boundary 20
- densification of height network 441
- densification of horizontal network 415
- densification of observed astrodreflections 573
- density
  - air 153
  - surface 486
- density function 486
- dependence
  - statistical 43
  - total 186
- deposit loading 132
- derivative
  - partial 31
  - total 31
- descending node 311
- design matrix 178
  - first 180
  - second 180
- design problem
  - combined 243
  - first-order 243
  - second-order 243
- design variance 227
- determinant of matrix 27
- deterministic model 183
- development of gravitational potential 470
- diagnosis of singularity 272
- diagonal matrix 26
  - n* 26
- diagonal of matrix 26
- diagonalization of matrix: eigenvalue 29
- diagram: commutative 38
- diffeomorphic transformation 355
- difference: range 318
- difference surface 574
- differential 31
  - total 31
- differential equation
  - ordinary 34
  - system of 35
  - partial 36
    - homogeneous 36
    - non-homogeneous 36
    - vector 35
- differential GPS positioning 343
- differential operator 32
- differential range 343
- differential rotation 655
- dilation 653
- direct problem on ellipsoid 352
- direct problem on mapping plane 363
- direct wave 155
- direction
  - east 293
  - horizontal 181
  - vertical 85
  - west 293
- direction and range mathematical model: simultaneous 317
- direction (to satellite) mathematical model 316
- direction to satellite 341
- Dirichlet boundary value problem 36
- dispersion 160
- displacement
  - coseismic 141
  - horizontal 639
  - vertical 615
    - absolute 619
    - relative 619
- displacement gradient matrix 649
- displacement model
  - constant acceleration 647
  - constrained
    - spatially 647
    - temporally 647
  - constrained magnitude 647
  - simple 644
  - slip 647

- spatially continuous 648
- displacement profile 619
- displacement vector: tidal 595
- distance(s) 37
  - correction to 352
  - difference of 181
  - horizontal 181
  - mean quadratic 195
  - satellite 181
  - spatial 181
  - variation of 181
    - tidal 597
  - zenith 294
- distortion
  - analytical model of 415
  - systematic network 439
- distribution parameter 42 *see also* population parameter
- disturbing potential 483
  - Molodenskij integral equation for 486
- diurnal aberration 301
- diurnal parallax 301
- Doodson tidal constant 589
- Doppler count (integrated) 320
  - correction to
    - ionospheric 322
    - tropospheric 321
- Doppler effect 318
- Doppler equation 318
- Doppler levelling 373
- Doppler principle 318
- drag-free satellites 548
- dummy variable 33
- Dupin indicatrix 41
- dyadic matrix 27
- dyadic product 32
- dynamic (height) correction 370
- dynamic flattening 65
- dynamic height 369
- dynamic model 277
- dynamic model error 278
- dynamic process 184
- dynamic geodesy 45
- dynamically undisturbed sea level 593
- dynamics
  - atmospheric 161
  - ocean 128
- E *see* ecliptical coordinate system
- earth's surface: continuous representation of 104
- east direction 293
- eccentric anomaly (orbital) 310
- ecliptic 58
- ecliptical coordinate system 302
- ecliptical latitude 302
- ecliptical longitude 302
- ecology 20
- economic parameter 201
- EDM *see* electronic distance measuring
- effect: Doppler 318
- eigenfunction 35
- eigenfunction series 35
- eigenvalue diagonalization of matrix 29
- eigenvalue of matrix 29
- eigenvector of matrix 29
- eikonal 157
- EL *see* ellipsoidal coordinate system (one parametric)
- elasticity of network 413
- electromagnetic force orbital perturbation 549
- electronic distance measuring (EDM) 181
- element(s)
  - Keplerian orbital 311
  - sequence of 30
  - series of 30
- element of a matrix 26
- ellipse *see also* confidence ellipse
  - orbital 310
- ellipsoid 40 *see also* confidence ellipsoid; reference ellipsoid
  - direct problem on 352
  - hydrostatic equilibrium 118
  - inverse problem on 354
- ellipsoidal angle 355
- ellipsoidal azimuth 350
- ellipsoidal harmonic function 476
- ellipsoidal model field 477
- ellipsoidal coordinate system
  - geodetic 38
  - one parametric (EL) 464
- elliptical integral 33
- elliptical point 41
- elongation 308
- energy
  - sink 460
  - source 460
- engineering project 20
- environmental management 20
- ephemeris 313
  - broadcast 315
  - precise 315
- ephemeris time 303
- equations(s) *see also* formula
  - Cauchy-Riemann 358
  - Clairaut 118
  - differential *see* differential equation
  - disturbing potential 483
  - Doppler 318

- Euler (free nutation)** 63  
**Euler (ordinary differential)** 469  
**Fredholm linear integral** 37  
**fundamental gravimetric** 495  
**homogeneous** 36  
**integral** 37  
**Kepler** 311  
**Laplace azimuth** 331  
**Laplace (barometric)** 443  
**Laplace (potential)** 36  
**Legendre** 35  
**Liouville** 66  
**linear: system of** 28  
**Lorenz-Lorentz** 156  
**Molodenskij** 486  
**normal** *see* **normal equations**  
**observation** *see* **observation equation**  
**ordinary differential** *see* **differential equation**  
**personal** 307  
**phase** 282  
**Poisson** 36  
**potential** 462  
**simultaneous linear** 28  
**summation** 282  
**vector differential** *see* **differential equations**  
**equation of continuity** 162  
**equation of motion**  
  **canonical** 550  
  **hydrodynamic** 162  
**equation of state** 154  
**equator** 293  
**equatorial orbit** 550  
**equinoctial colure** 294  
**equipotential surface** 84  
**erosion rebound** 137  
**error**  
  **dynamic model** 278  
  **random** 197  
  **systematic** 198  
**error ellipsoid (ellipse)** *see* **confidence ellipsoid (ellipse)**  
**error propagation law** 197  
  **power** 435  
  **square root** 433  
**escarpment** 143  
**establishment of horizontal geodetic datum** 327  
**estimate of unknown parameter** 205  
**estimated value of observable** 188  
**estimation** 44  
**Euclidean metric** 195  
**Euler (free nutation) equation** 63  
**Euler (ordinary differential) equation** 469  
**Euler formula** 41  
**Euler period** 65  
**eustatic sea level variation** 148  
**evaporation rebound** 137  
**exact value of observable** 188  
**expansion**  
  **point of** 33  
  **Taylor** 33  
**expansion factor** 230  
**expansion of horizontal network** 414  
**expectation operator** 43  
**expected residual** 202  
**expected value of observable** 183  
**experimental probability** 42  
**exploration geophysics** 21  
**extension** 652  
**external gravitational potential** 463  
**extremum**  
  **absolute** 31  
  **local** 31  
**factor**  
  **expansion** 230  
  **point scale** 356  
  **variance** 211  
    **a posteriori** 213  
**family of coordinate systems** 38  
**faulting** 143  
**Fédération Internationale des Géomètres** 49  
**Fermat principle** 157  
**field**  
  **conservative** 82  
  **ellipsoidal** 477  
  **gravity** *see* **gravity field**  
  **irrotational** 82  
  **radial** 477  
  **scalar** 36  
  **vector** 36  
**FIG** *see* **Fédération Internationale des Géomètres**  
**filter** 249  
  **band-pass** 257  
  **Bayes** 281  
  **convolutive** 249  
  **covariance** 261  
  **high-pass** 257  
  **Kalman** 277  
  **linear** 249  
  **low-pass** 257  
  **normalized** 251  
  **predictive** 250  
  **recursive** 249  
  **response of** 251  
  **sequential** 250  
    **symmetrical** 251  
**filtering** 249  
**first design matrix** 180

- first law of thermodynamics 162  
 first Love number 590  
 first-order atmospheric correction to gravity 164  
 first-order atmospheric correction to gravity potential 164  
 first-order design problem 243  
 fixed object 416  
 flat spectrum 256  
 flattening  
     dynamic 65  
     gravity 114  
     hydrostatic 119  
 floating datum 399  
 floating object 416  
 fluctuation: earth's spin velocity 68  
 folding frequency 254  
 footprint of altimeter 444  
 forbidden spherical harmonic 518  
 force  
     centrifugal 73  
     gravitational 71  
     gravity 74  
     perturbing 552  
     tidal 124  
 forced nutation 61  
 form: quadratic 27  
 formula *see also* equation  
     Bruns 493  
     generalized 493  
     Euler 41  
     Frenet 39  
     Gauss 34  
     Gauss mid-latitude 354  
     International Gravity  
         1930 78  
         1967 79  
         1980 78  
     Legende (decomposition) 473  
     long line 353  
     Puissant 353  
     short line 353  
     Somigliana 482  
     Stokes 520  
     Vening Meinesz 521  
     Wiener-Kolmogorov 264  
 four-dimensional model for vertical displacement 630  
 Fourier-method 36  
 Fourier series  
     generalized 36  
     trigonometrical 36  
 Fourier spectrum 254  
 Fourier transformation 254  
     complex 255  
     Fredholm linear integral equation 37  
     free-air gravity anomaly 79  
     free-air gravity correction 77  
     free-air gravity gradient 499  
     free-air geoid 523  
     free-fall device 534  
     free nutation 63  
     free oscillation 147  
     Frenet formula 39  
     frequency  
         folding 254  
         fringe 344  
         Nyquist 254  
         resonant 560  
         tidal 127  
         frequency offset 320  
     frequency space 253  
     fringe frequency (in VLBI) 344  
     function  
         autocovariance 186  
         base 246  
         complex 30  
         constraint 180  
         covariance 186  
         cross-covariance 189  
         generating 33  
         Green 37  
         Hamiltonian 550  
         harmonic *see* harmonic function  
         Legendre 33  
             second kind 476  
         Legendre associated 468  
         likelihood 268  
         matrix 30  
         one valued 30  
         probability density *see* probability density function  
         real 30  
         Stokes 519  
             spheroidal 578  
         surface density 486  
         transition 184  
         transition probability 185  
         variation 205  
         vector 30  
         Vening Meinesz 522  
     function of several (scalar) variables 31  
     function of several vector variables 31  
     functional value 31  
     fundamental gravimetric equation 495  
     fundamental harmonic function 463  
     fundamental quantity: Gaussian 41  
     fundamental rotation matrix 38

- G *see* geodetic coordinate system  
 gain matrix 280  
*GAST* *see* Greenwich apparent sidereal time  
 Gauss formula 34  
     mid-latitude 354  
 Gaussian fundamental quantity 41  
 generalized adjustment 265  
 generalized Bruns formula 493  
 generalized coordinate 549  
 generalized Fourier series 36  
 generalized matrix inverse *see also* g-inverse  
     least squares 275  
     minimum norm 275  
     Moore-Penrose 275  
 generalized momentum 549  
 generating function 33  
 geocentric coordinate system 296  
 geocentric datum position parameter *see* datum  
     (position) parameter  
 geocentric reference ellipsoid 88, 115  
     biaxial 110  
     international 114  
     mean earth's 87  
     triaxial 107  
 geodesic (curve) 40  
 geodesy  
     dynamic 45  
     geometrical 45  
     International Association of 47  
     mathematical 45  
     physical 45  
 geodetic astronomy 292  
 geodetic azimuth 328  
 geodetic base line 637  
 geodetic boundary value problem 517  
 geodetic coordinate system 324  
     ellipsoidal 38  
     local 328  
 geodetic height 365  
 geodetic height network *see* height network  
 geodetic horizontal network *see* horizontal network  
 geodetic latitude 38  
 geodetic levelling 366  
 geodetic longitude 38  
 geodetic meridian 328  
 geodetic methodology 175  
 geodetic north 328  
 geodetic parameter 119  
 geodetic reference ellipsoid *see* reference ellipsoid  
 geodetic scientist 50  
 geodetic vertical angle *see* vertical angle  
 geodetic zenith distance *see* zenith distance  
 geography 20  
 geoid 87  
     astro-geodetic 566  
     astro-gravimetric 577  
     free-air 523  
     instantaneous 617  
     isostatically compensated 524  
 geoid matching 570  
 geoidal deflection *see* deflection of vertical  
 geoidal gravity anomaly *see* gravity anomaly  
 geoidal height (undulation) 89  
     absolute 89  
     relative 89  
     tidal variation 593  
 geology 23  
 geometrical attenuation factor 560  
 geometrical geodesy 45  
 geometrical mode of simultaneous positioning (by  
     satellites) 385  
 geometrical space 37  
 geophysics 21  
     exploration 21  
 geopotential number 368  
 geostrophic wind 162  
 geosyncline 143  
 g-inverse *see* generalized matrix inverse  
 glacial (ice) melt effect on sea level variation 426  
 global circulation 427  
 global gravity network 535  
 Global Positioning System (GPS/NAVSTAR)  
     314  
 goodness of fit test 227  
 GPS positioning 314  
     differential 343  
 graben 143  
 gradient matrix: displacement 649  
 gradient of gravity *see* gravity gradient  
 gradiometry: satellite 563  
 Gram matrix 247  
 Gram-Schmidt process 248  
 gravimeter 534  
 gravimetric deflection *see* deflection of vertical  
 gravimetric equation: fundamental 495  
 gravitational attraction 71  
 gravitational attraction of tidal water 600  
 gravitational constant 71  
 gravitational flux 460  
 gravitational force 71  
 gravitational potential 83  
     external 463  
     internal 463  
 gravitational wave 147  
 gravity 181  
     gradient of *see* gravity gradient  
     normal 78

- reference 77
- tidal 127
- topographical effect on 509
- gravity anomaly** 79
  - Bouguer 513
  - free-air 79
  - geoidal 489
  - isostatically compensated 514
  - mean 537
    - prediction of 538
  - point 536
  - surface 490
- gravity correction (to height) 530
- gravity difference 181
- gravity equipotential surface *see* equipotential surface
- gravity field** 75
  - earth's 46
  - effect of isostasy on 514
  - ellipsoidal 477
  - model 477
  - normal 78
  - parameter of 544
  - radial 477
    - temporal variation of 46
- gravity flattening 114
- gravity force 74
- gravity formula**
  - International 1930 78
  - International 1967 79
  - International 1980 79
- gravity gradient (vertical) 181
  - Bouguer 502
  - Bouguer plate 502
  - curvature 502
  - free-air 499
  - normal 498
  - Poincaré-Pray 500
  - topographical 513
  - total surface 513
- gravity measurement
  - absolute 534
  - relative 534
  - repeated 614
- gravity network
  - global 535
  - national 535
- gravity potential 83
- gravity standardization net: international (IGSN71) 535
- gravity system: Potsdam 535
- gravity variation due to oblateness of earth 77
- gravity variation due to polar motion 608
- gravity variation due to tide 127
- gravity variation with height 77
- gravity vector 75
  - normal 78
- Green function 37
- Green method 37
- Green second identity 34
- Greenwich apparent sidereal time 298
- Greenwich observatory
  - instantaneous 298
  - mean 296
- grid azimuth 362
- grid north 362
- ground compaction 143
- ground swelling 146
- ground wave 155
- group velocity 160
- gyroscope 59, 339
- Hamiltonian function** 550
- harmonic analysis 252
- harmonic function** 36
  - ellipsoidal 476
  - fundamental 463
  - spherical *see* spherical harmonic (function)
- harmonic motion: simple 35
- height**
  - barometric 443
  - difference of 181
  - dynamic 369
  - geodetic 365
  - geoidal 89
    - absolute 89
    - relative 89
  - gravity correction to 430
  - normal 372
  - orthometric 370
    - Helmert 371
    - Vignal 373
  - tidal variation of 592
  - trigonometrical 364
- height anomaly 118
- height difference based on normal gravity 373
- height difference determined by three-dimensional methods 373
- height network (geodetic)** 97
  - adjustment of 429
  - densification of 441
  - design of 439
  - distortion of 439
  - kinematical adjustment of 625
  - origin of 428
- height of target correction 350
- heliocentric coordinate system 293
- Helmert blocking** 405

- Helmert orthometric height 371  
 Hessian matrix 211  
 high orbiting satellite 546  
 high-pass filter 257  
 Hilbert space 196  
 Hilbert space optimization 196  
 histogram 41  
 homogeneous partial differential equation 36  
 homogeneous kernel 31  
 horizon: celestial 293  
 horizontal angle 181  
     ellipsoidal 355  
     plane 356  
     projected 356  
     tidal variation of 597  
 horizontal angle correction (terrain to ellipsoid) 351  
 horizontal angle observation equation 377, 404  
 horizontal angle to star 181  
 horizontal datum positioning *see* datum positioning  
 horizontal direction 181  
 horizontal direction correction (terrain to ellipsoid) 350  
 horizontal displacement 639  
 horizontal distance 181  
 horizontal (geodetic) datum 99  
     establishment of 327  
     floating 399  
     parameter of *see* datum parameter  
     positioning of *see* datum positioning  
 horizontal network (geodetic) 99  
     abstract elasticity of 413  
     adjustment of 401  
     densification of 415  
     expansion of 414  
     merger of 416  
     optimum design analysis of 409  
     origin of 380  
     resurvey of 634  
     simulated 412  
     strain of 414  
     strength of 409  
     systematic distortions of 413  
 horizontal network in three dimensions 401  
 horizontal network on conformal mapping plane 403  
 horizontal network on reference ellipsoid 401  
 horizontal position: comparison of 638  
 horizontal relative positioning on conformal map §16.3  
 horizontal relative positioning on reference ellipsoid §16.2  
 horizontal refraction 158  
 horizontal strain  
     polar motion 608  
     tidal 595  
 hour angle 302  
 Householder orthogonal transformation 208  
 humidity: air 155  
 hydrodynamic equation of motion 162  
 hydrographic chart 441  
 hydrographic chart datum 441  
 hydrography 21  
 hydrostatic equilibrium ellipsoid 118  
 hydrostatic flattening 119  
 hyperbolic point 41  
 hyperbolic positioning 319, 419  
 hyperellipsoid 29  
 hypermatrix 27  
 hypothesis *see* statistical hypothesis  
 IAG *see* International Association of Geodesy  
 IAU *see* International Astronomical Union  
 ICA *see* International Cartographic Association  
 ice loading 131  
 ice (glacial) melt loading 131  
 identity: Green's second 34  
 IGSN 71 *see* International Gravity Standardization Net 1971  
 ill-conditioned matrix 27  
 ILS *see* International Latitude Service  
 imaginary part of a complex number 29  
 implicit model 180  
     linear 180  
 in-context (simultaneous) point confidence interval 439  
 in-context test 229  
 inclination 311  
 incomplete gravimetric deflection *see* deflection of vertical  
 independence: statistical 43  
 indeterminable parameter 272  
 indeterminacy 272  
 indeterminacy in rotation 640  
 indeterminacy in scale 640  
 indeterminacy in translation 639  
 index of refraction 156  
 indicatrix  
     Dupin 41  
     Tissot 41  
 indirect effect of tidal water 600  
 indirect effect (of  $\Delta g$ ) on computed geoid 524  
 indirect effect (of  $\Delta g$ ) on disturbing potential 524  
 inequality: Bonferroni 231  
 inertia: tensor of 63  
 inertial positioning 337

- inertial coordinate system 302  
 initial point of network 380  
 initial value 35  
 initial value problem 35  
 inner constraint 274  
 inner (scalar) product 35  
 innermost block 541  
 instantaneous astronomical azimuth 298  
 instantaneous geoid 617  
 instantaneous Greenwich observatory 298  
 instantaneous orthometric height 606  
 instantaneous sea level 424  
 instantaneous sea surface topography 427  
 instantaneous spin axis 59  
 instantaneous terrestrial coordinate system 297  
 instrument 177  
 integral
  - angular-momentum 547
  - convolution 33
  - elliptical 33
  - line 33
  - Riemann 33
  - singularity of 535
  - Stokes 519
  - surface 34
  - volume 34
 integral equation 37
  - Fredholm linear 37
  - Molodenskij 486
 integral of matrix function 33  
 integral of vector function 33  
 integrated Doppler count 320  
 integration constant 35  
 interference: non-linear 128  
 internal gravitational potential 463  
 International Association of Geodesy 47  
 International Astronomical Union 66  
 International Cartographic Association 50  
 international (reference) ellipsoid 114  
 International Federation of Surveyors 49  
 international gravity formula
  - 1930 78
  - 1967 79
  - 1980 79
 International Gravity Standardization Net 1971
  - 535
 International Latitude Service 66  
 International Polar Motion Service 66  
 International Society for Photogrammetry 50  
 International Union of Geodesy and Geophysics
  - 47
 interpolation 246  
 interstation vector 336  
 interval: confidence *see* confidence interval  
 inverse
  - matrix 28
  - generalized (g-inverse) 275 *see also* generalized matrix inverse
 inverse azimuth 354  
 inverse problem on ellipsoid 354  
 inverse problem on mapping plane 363  
 inverse problem of relative three-dimensional positioning 345  
 ionosphere 152  
 ionospheric correction to Doppler count 322  
 ionospheric delay correction 315  
 ionospheric refraction 160  
 IPMS *see* International Polar Motion Service  
 irregularities (of gravitational field)
  - tesseral 549
  - zonal 549
 irrotational field 82  
 isobaric surface 154  
 isobars 154  
 isometric latitude 359  
 isometric plane 360  
 isostasy 133  
 isostatic rebound 136
  - postglacial 136
 isostatically compensated geoid 524  
 isostatically compensated gravity anomaly 514  
 isotropic kernel 31  
 ISP *see* International Society for Photogrammetry  
 IT *see* instantaneous terrestrial coordinate system  
 IUGG *see* International Union of Geodesy and Geophysics  
 Jacobian 34  
 Jacobian matrix 26  
 Kalman filter 277  
 Kepler equation 311  
 Keplerian motion 310  
 Keplerian orbital element 311  
 kernel 30
  - homogeneous 31
  - isotropic 31
  - weight 543
 kinematical adjustment of height network 625  
 kinematical loop misclosure 622  
 kinematically constrained mode of satellite positioning 387  
 $L_2$  space 196  
 LA *see* local astronomical coordinate system  
 Lagrange correlate 205  
 Lagrange method 206

- lake level variation 613
- landslide deformation 146
- Laplace azimuth 348
- Laplace barometric equation 443
- Laplace correction 348
- Laplace (potential) equation 36
- Laplace (azimuth) equation 331
- large city loading 133
- laser ranging
  - lunar 314
  - satellite 314
- LAST *see* local apparent sidereal time
- latitude
  - astronomical 296
  - ecliptical 302
  - geodetic 38
  - isometric 359
- latitude determination by meridian zenith distances 306
- latitude mathematical model 305
- law
  - covariance 197
  - error propagation 197
    - power 435
    - square root 433
  - Snell 138
  - universal gravitation 70
- LBI *see* long base line interferometry
- least-squares collocation 261
- least-squares estimate of observation 207
- least-squares estimate of unknown parameter 205
- least-squares g-inverse 275
- least-squares residual 207
- least-squares solution 200, 205
  - properties of 219
- least-squares spectrum 256
- Legendre associated function 468
- Legendre decomposition formula 473
- Legendre equation 35
- Legendre function 33
- Legendre function of second kind 476
- level
  - confidence 221
    - critical 223
  - significance 221
    - critical 223
- level surface 85
- levelled height difference
  - loading correction to 606
  - tidal variation of 598
- levelling
  - astro-geodetic 566
  - Doppler 373
  - geodetic 366
- refraction correction in 431
- steric 427
- levelling line 430
- levelling loop misclosure 433
  - kinematical 622
  - static 433
- levelling network 429
- levelling segment 433
- LG *see* local geodetic coordinate system
- likelihood: maximum 258
- likelihood function 268
- limit 30
- line
  - levelling 430
  - nodal 61
  - plumb 84
- line integral 33
- line of apsides 310
- line scale 362
- line spectrum 256
- linear condition model 179
- linear equation 28
- linear filter 249
- linear form 245
- linear implicit model 180
- linear integral equation: Fredholm 37
- linear model explicit in  $t$  180
- linear model explicit in  $x$  178
- linear operator 26
- linear orbital perturbation 559
- linear (vector) space 25
- linear transformation 39
- linear vertical movement 622
- Liouville equation 66
- lithosphere 130
  - continental 134
  - density of 134
  - oceanic 134
- load number 600
- loading
  - barometric pressure 133
  - deposit 132
  - ice 131
  - ice melt 131
  - large city 133
  - magma 133
  - precipitation 133
  - snow 133
  - tidal water 132
  - water reservoir 132
- loading correction to levelled height difference 606
- loading correction to orthometric height 606
- local apparent sidereal time 302

- local astronomical coordinate system (LA) 294  
 local extremum 31  
 local geodetic coordinate system (LG) 328  
 local instantaneous sea level 424  
 local mean sea level 424  
 local (sea level) response technique 428  
 locally compact space 30  
 locally linear transformation 39  
 locally orthonormal coordinate system 39  
 long-arc prediction 554  
 long arc satellite positioning 388  
 long base line interferometry 344  
 long-line formula 353  
 long orbital arc 388  
 long periodic orbital variation 559  
 long periodic tidal effect on sea level variation 426  
 longitude  
     astronomical 296  
     ecliptical 302  
     geodetic 38  
     nutation in 300  
 longitude determination by transit times 308  
 longitude determination by zenith distances 307  
 longitude mathematical model 307  
 loop: closed 368  
 Lorenz-Lorentz equation 156  
 Love number  
     first 590  
     second 592  
     third (Shida) 592  
 low-pass filter 257  
 lower triangular matrix 27  
 lunar laser ranging 314  
 luni-solar precession 62  
 luni-solar tidal potential 126
- M** *see* mapping coordinate system  
 magma loading 133  
 magnetosphere 153  
 main diagonal of matrix 27  
 main system of normal equations 406  
 man made deformations §8.4  
 management  
     environmental 20  
     urban 20  
 mapping 20  
     cartographical 334  
     conformal 357  
 mapping coordinate system 334  
 mapping plane 363  
 marine positioning 416  
 Markov chain 185  
 mascons 487
- matching: geoid 570  
 mathematical geodesy 45  
 mathematical model 176  
     astronomical azimuth 308  
     astronomical latitude 305  
     auxiliary 183  
     combination of 181  
     condition 179  
     constrained displacement magnitude 647  
     deterministic 183  
     direction (to satellite) 316  
     displacement  
         constant acceleration 647  
         simple 644  
         slip 647  
         spatially constrained 647  
         spatially continuous 648  
         temporally constrained 647  
         vertical 630  
     dynamic 277  
     hyperbolic positioning 419  
     implicit 180  
     intersection 418  
     latitude 305  
     linear condition 179  
     linear implicit 180  
     longitude 307  
     primary 181  
     range 313  
     range difference 321  
     reformulation of 200  
     resection 419  
     secondary 181  
     simultaneous direction and range 317  
     singular 272  
     stochastical 183  
     mathematical model explicit in  $\lambda$  179  
         linear 180  
     mathematical model explicit in  $x$  178  
         linear 178  
     mathematical model for horizontal network on conformal mapping plane 403  
     mathematical model for horizontal network on reference ellipsoid 401  
     mathematical model for horizontal network in three dimensions 401  
     mathematical model for kinematical adjustment of height network 625  
     mathematical model for relative positioning by differential ranges 343  
     mathematical model of distortions 415  
     mathematical statistics 214  
     mathematics 23  
     matrix 25

- antisymmetrical 26
- bivariate covariance 232
- coefficient 205
- covariance 43
- cross-covariance 211
- defect of 27
- design 178
- determinant of 27
- diagonal 26
- diagonal of 27
- diagonalization of 29
- displacement gradient 649
- dyadic 27
- eigenvalue diagonalization of 29
- eigenvalue of 29
- eigenvectors of 29
- elements of 26
- first design 180
- fundamental rotation 38
- gain 280
- generalized inverse of *see* generalized matrix inverse
- Gram 247
- Hessian 211
- ill-conditioned 27
- inverse of 28
- Jacobian 26
- lower triangular 27
- main diagonal of 27
- n*-diagonal 26
- orthogonal 27
- partitioning of 27
- positive definite 27
- profile of 407
- rank deficient 28
- rank of 28
- reflection 38
- regular 27
- rotation 38
- second design 180
- singular 27
- square 26
- symmetrical 26
- trace of 27
- transition 184
- unit 27
- upper triangular 27
- Vandermonde 26
- weight 211
- matrix function 30
  - integral of 33
- matrix profile 407
- maximum cutoff spectral analysis 258
- maximum likelihood 258
- McLaurin series 33
- mean
  - bivariate population 232
  - normal test on 227
  - probability density function 42
    - multidimensional 43
  - sample 42
    - Student t test on 227
- mean (orbital) anomaly 310
- mean earth's ellipsoid 87
- mean gravity anomaly 537
- mean Greenwich Observatory 296
- mean quadratic approximation 247
- mean quadratic distance 195
- mean right ascension coordinate system (A) 294
- mean sea level 105
  - local 424
- mean value theorem 33
- measurement 177
- merger of horizontal networks 416
- merger of three-dimensional networks 393
- meridian
  - astronomical 293
  - geodetic 328
- meridian component of deflection of vertical 93
- meridian convergence 360
- meridian radius of curvature 111
- mesopause 152
- mesosphere 152
- method
  - Cholesky 209
  - Fourier 36
  - Green 37
  - Lagrange 206
- method of azimuth determination by hour angles of stars near culmination 308
- method of azimuth determination by hour angles of stars near elongation 309
- method of latitude determination by meridian zenith distances 306
- method of longitude determination by transit times 308
- method of longitude determination by zenith distances 307
- method of separation of variables 36
- methodology: geodetic 175
- metric 194
  - axiom of 194
  - Euclidean 195
- metric space 194
- metric tensor 196
- mid-latitude formula: Gauss 354
- minimal constraint 273
- minimum constraint solution 273

- minimum distance condition 194  
 minimum norm condition 194  
 minimum norm g-inverse 275  
 minimum quadratic form of weighted residuals 201  
 misalignment angle: datum 327  
 misclosure  
     kinematical (levelling) loop 622  
     quadratic form of 237  
     statical (levelling) loop 433  
 misclosure vector 203  
 mode of satellite positioning *see* satellite positioning  
 model *see also* mathematical model  
 model error: dynamic 278  
 model gravity field 477  
 model space 178  
 Molodenskij deflection of vertical 491  
 Molodenskij integral equation for disturbing potential 486  
 Molodenskij solution of geodetic boundary value problem §22.2  
 momentum 549  
 monitoring network 636  
 Moore-Penrose g-inverse 275  
 most powerful test 222  
 motion  
     canonical equation of 550  
     hydrodynamic equation of 162  
     Keplerian 310  
     polar 66  
     proper 299  
     simple harmonic 35  
 moving average 251  
 MRA *see* mean right ascension coordinate system  
 MSL *see* mean sea level  
 multidimensional population parameter 219  
 multivariate  
     random 42  
     stochastical 42  
 multivariate analysis 232  
 multivariate subset analysis 232  
 multivariate test 233  
  
 NAD *see* North American datum  
 nadir 293  
 national gravity network 535  
 navigation 416  
 NAVSTAR Global Positioning System 314  
 NCP *see* north celestial pole  
 $n$ -diagonal matrix 26  
 neighbourhood of point: compact 30  
 network
- geodetic 97  
 gravity  
     global 535  
     national 535  
 height *see* height network  
 horizontal *see* horizontal network  
 levelling *see* levelling network  
 monitoring 636  
 origin of 99, 380  
 photogrammetrical §17.2  
 resurveyed horizontal 634  
 three-dimensional *see* three-dimensional network  
 Neumann boundary value problem 36  
 nodal line 61  
 node  
     ascending 311  
     descending 311  
 noise 184  
     random 256  
     systematic 256  
     white 256  
 non-homogeneous partial differential equation 36  
 non-Keplerian orbit 548  
 non-linear form of normal equations 208  
 non-linear interference in ocean dynamics 128  
 non-linear solution 192  
 non-linearity: effect of 207  
 non-parametric coordinate system 38  
 non-parametric statistics 215  
 norm 35  
     quadratic 195  
 normal body of earth 87  
 normal (height) correction 429  
 normal equations  
     non-linear form of 208  
     phase 282  
     sequential 281  
     summation 282  
     system of 204  
         main 406  
 normal gravity field 78  
 normal gravity gradient 498  
 normal gravity vector 78  
 normal height 372  
     tidal variation of 593  
 normal plane 39  
 normal plumb line 491  
 normal potential 88  
 normal potential coefficient 480  
 normal probability density function 42  
     bivariate 232  
 normal section to geodesic correction 350  
 normal test on mean 227

- normal vector 39  
 normalized filter 251  
 normalized spherical harmonic function 469  
 normed space 194  
 north  
     geodetic 328  
     grid 362  
 North American datum 89  
 north celestial pole 292  
 nuisance parameter 183  
 null hypothesis 220  
 number  
     complex 29  
     first Love 590  
     geopotential 368  
     load 600  
     real 25  
     second Love 592  
     Shida 591  
     third Love 592  
 nutation 61  
     forced 61  
     free 63  
 nutation in longitude 300  
 nutation in obliquity 300  
 Nyquist frequency 254
- object  
     anchored 416  
     fixed 416  
     floating 416  
 oblateness of earth 77  
 obliquity 60, 300  
     nutation in 300  
 observable 177  
     decomposition of 187  
     estimated value of 188  
     exact value of 188  
     expected value of 183  
     representative value of 182  
     systematic component of 183  
 observation 177  
     least-squares estimate of 207  
     quasi 193  
     table of 182  
 observation equation for horizontal network in three dimensions  
     astronomical azimuth 401  
     direction 401  
     horizontal angle 401  
     spatial distance 401  
 observation equation for horizontal network on conformal mapping plane  
     chord distance 403  
     direction 404  
     grid azimuth 404  
     horizontal angle 404  
 observation equation for horizontal network on reference ellipsoid  
     direction 403  
     ellipsoidal distance 402  
     geodetic azimuth 402  
     horizontal angle 403  
 observation equation for levelling line 432  
 observation equation for point velocity 626  
 observation equation for relevelled segment 626  
 observation equation for three-dimensional network  
     astronomical azimuth 375  
     astronomical coordinate 378  
     deflection component 379  
     direction 377  
     height difference 379  
     horizontal angle 377  
     spatial distance 378  
     vertical angle 377  
 observation in space 182  
 observation in time 182  
 observation space 178  
 observed terrain tilt 598  
 observing list 307  
 ocean dynamics 128  
 ocean global circulation 427  
 oceanic block 135  
 oceanic lithosphere 134  
 oceanography 22  
 offset  
     frequency 320  
     tracking station clock 314  
 (1- $\alpha$ ) confidence interval 224  
 one parametric ellipsoidal system of coordinates 464  
 one-sided probability values 223  
 one-tailed test 224  
 one-valued function 30  
 operator  
     differential 32  
     expectation 43  
     linear 26  
 optimization in Hilbert space 196  
 optimum design analysis of horizontal network 409  
 OR see orbital coordinate system  
 orbit  
     circular 550  
     equatorial 550  
     non-Keplerian 548  
     prediction of 554

- secular variation of 559
- short periodic variation of 559
- orbital arc** 319
  - long 388
  - short 387
- orbital coordinate system** 311
- orbital element** 311
  - Keplerian 311
- orbital ellipse** 310
- orbital perturbation** 554
  - electromagnetic force 549
  - linear 559
  - secular 557
  - tidal 549
- orbital variation**
  - long periodic 559
  - secular 559
  - short periodic 559
- orbiting satellite**
  - close 546
  - high 546
- ordinary differential equation** 34
  - system of 35
- orientation of a coordinate system** 37
- orientation unknown** 377
- origin of a coordinate system** 37
- origin of a height network** 428
- origin of a horizontal network** 380
- orthogonal base** 247
- orthogonal decomposition** 199
- orthogonal matrix** 27
- orthogonal transformation: Householder** 208
- orthogonality of eigenvector** 29
- orthometric (height) correction** 429
- orthometric height** 370
  - Helmert 371
  - loading correction to 606
  - tidal variation of 593
  - Vignal 373
- orthonormal base** 248
- orthonormal system** 39
- oscillation: free** 147
- oscillatory phenomenon** 250
- osculating plane** 39
- out-of-context point confidence ellipse** 355
- out-of-context point confidence ellipsoid** 347
- out-of-context point confidence interval** 439
- out-of-context test** 229
- outer block** 543
- outlier** 227
- overconstrained solution** 273
- overdetermined solution** 193
- parallax**
- annual 300
- diurnal 301
- parallelism: conditions for** 331
- parameter**
  - adjusted 205
  - datum 327
  - distribution 42
  - economic 201
  - geocentric datum position 327
  - geodetic 119
  - indeterminable 272
  - least-squares estimate of 205
  - nuisance 183
  - population 218
    - multidimensional 219
  - topocentric datum position 330
  - unknown 177
- parameter space** 178
- parameter vector** 188
- parametric statistics** 214
- partial derivative** 31
- partial differential equation** 36
  - unhomogeneous 36
- particular solution** 35
- particular solution for position** 639
- partitioning of matrix** 27
- pass: satellite** 319
- PCA** see *point of closest approach*
- pear-shaped reference body** 109
- pedal curve** 357
- pendulum** 534
- perigee** 310
  - argument of 311
- perihelion** 58
- period**
  - Chandler 66
  - Euler 65
  - tidal 127
- periodogramme** 254
- permanent tidal uplift** 128
- personal equation** 307
- perspective centre** 381
- perturbing forces** 552
- perturbing potential** 548
- phase equation** 282
- phase lag of surface wave** 422
  - primary 421
  - secondary 421
- phase of trigonometrical term** 253
- phase of response** 251
- photogrammetrical block** 381
- photography: two-media** 442
- physical geodesy** 45
- physics** 23

- plane  
 isometric 360  
 mapping 363  
 normal 39  
 osculating 39  
 rectifying 39  
 plane angle 356  
 planetary precession 302  
 planetology 21  
 plumb line 84  
 normal 491  
 Poincaré-Pray gradient 500  
 point  
 circular 41  
 crossover 445  
 deflection 565  
 elliptical 41  
 expansion 33, 193  
 hyperbolic 41  
 initial 380  
 neighbourhood of 30  
 prediction 248  
 regular 41  
 sample 182  
 vernal 61  
 point confidence ellipse 355  
 in-context 408  
 out-of-context 408  
 standard 355  
 point confidence ellipsoid 346  
 in-context 392  
 out-of-context 390  
 standard 347  
 point confidence interval  
 in-context (simultaneous) 439  
 out-of-context 439  
 point gravity anomaly 536  
 point load 601  
 point of closest approach (PCA) 318  
 point positioning Ch. 15  
 repeated 634  
 satellite 310  
 point scale factor 356  
 point tilt variation 614  
 point velocity 626  
 Poisson equation 36  
 polar coordinate system on sphere 540  
 polar motion (wobble) 66  
 polar motion deformation 147  
 polar motion gravity variance 608  
 polar motion horizontal strain 608  
 polar motion potential 607  
 polar motion stress 609  
 polar motion uplift 607  
 polar motion variation of tilt 608  
 polar wobble (motion) 66  
 polarity of coordinate system 37  
 pole  
 north celestial 292  
 south celestial 292  
 polygon 41  
 polynomial : trigonometrical 252  
 population mean: bivariate 232  
 population parameter 218, *see also* probability distribution parameters  
 multidimensional 219  
 position(s)  
 comparison of horizontal 638  
 repeated vertical 612  
 temporal variation of 46  
 position vector 37  
 positioning  
 astronomical §15.2  
 directions to satellites 341  
 Doppler (range differences) 318  
 GPS 314  
 GPS differential 343  
 horizontal Ch. 15  
 horizontal datum 396  
 hyperbolic 319, 419  
 inertial 337  
 marine 416  
 point Ch. 15  
 repeated 634  
 satellite 310  
 range differences (Doppler) 318  
 range-range 419  
 ranging to satellites 343  
 relative Ch. 16  
 satellite *see* satellite positioning  
 three-dimensional *see* three-dimensional positioning  
 TRANSIT 318  
 vertical Ch. 15  
 positioning system: self contained 422  
 positive definite matrix 27  
 postglacial isostatic rebound 136  
 potential  
 centrifugal 83  
 deformation 592  
 disturbing 483  
 external (gravitational) 259  
 gravitational 83  
 gravity 83  
 internal (gravitational) 259  
 luni-solar tidal 126  
 normal 88  
 perturbing 548

- polar motion 607  
 sea tide 601  
 spin velocity variation 609  
 tidal 125  
 potential coefficient 471  
     normal 480  
     satellite 562  
 potential equation 462  
     Laplace 36  
 Potsdam Gravity System 535  
 power law 435  
 power of test 222  
 power spectrum 258  
 preanalysis 176  
 precession 59  
     luni-solar 62  
     planetary 302  
 precessional constant 299  
 precipitation loading 133  
 precise ephemeris 315  
 precision 42  
 prediction 248  
 prediction of deflection of vertical 573  
 prediction of mean gravity anomaly 538  
 prediction of orbit 554  
     long-arc 554  
     short-arc 554  
 prediction of vertical crustal velocities 627  
 prediction point 248  
 prediction space 261  
 predictive filter 250  
 pressure: atmospheric (air) 153, 425  
 pressure gradient acceleration 162  
 primary model 181  
 primary phase lag 421  
 prime vertical 283  
 prime vertical component (of deflection of vertical) 93  
 prime vertical radius of curvature 324  
 principal axis of strain 653  
 principle: Fermat 157  
 probability 42  
     conditional 44  
     experimental 42  
     simultaneous 44  
 probability density function 42  
     bivariate normal 232  
     chi-squared 42  
     F Table 13.4  
     mean of 42  
     mean of multidimensional 43  
     normal 42  
     parameter of 42  
     standard normal Table 13.2  
     Student t Table 13.2  
     tau Table 13.3  
     uniform 42  
     variance of 42  
 probability measure 215  
 probability space 215  
 probability statement 42  
 probability value  
     one-sided 223  
     two-sided 224  
 problem  
     approximation 247  
     boundary value *see* boundary value problem  
     initial value 35  
     singularity 272  
 process  
     dynamic 184  
     Gram-Schmidt 248  
 product  
     dyadic 32  
     inner (scalar) 35  
     scalar (inner) 32, 35  
     vector 32  
 profile  
     displacement 619  
     matrix 407  
     spatial 625  
     temporal 625  
 projected angle 356  
 projected azimuth 361  
 projected geodesic 355  
 projects: engineering 20  
 propagation delay correction 315  
 propagation of random error 197  
 propagation of systematic error 198  
 proper motion of star 299  
 properties of least-squares solution 219  
 Puissant formula 353  
 quadratic approximation: mean 247  
 quadratic distance: mean 195  
 quadratic form 27  
     quadratic form of mis closures 237  
     quadratic form of residuals 237  
     quadratic form of weighted residuals 201  
     quadratic norm 195  
     quadric: central 29  
     quantities: Gaussian fundamental 41  
     quasigeoid 117  
     quasi-observation 193  
 RA *see* right ascension coordinate system  
 radial field 477  
 radio-interferometry 344

- radio ranging: satellite 314
- radius of curvature 40
  - meridian 111
  - prime vertical 324
- radius vector 37
- random error 197
- random multivariate 42
- random noise 256
- random sample 41
- random series 264
- random variable 42
- random walk theory 439
- range 313
  - differential 343
  - positioning by 313
- range and direction mathematical model: simultaneous 317
- range difference 318
  - positioning by 342
- range-range positioning 419
- range refraction correction 315
- ranging
  - lunar laser 314
  - satellite laser 314
  - satellite radio 314
- rank deficient matrix 28
- rank of matrix 28
- real function 30
- real number 25
- real part of complex number 29
- rebound
  - erosion 137
  - evaporation 137
  - isostatic 136
  - postglacial isostatic 136
- receiver: satellite 314
- rectifying plane 39
- recursive filter 249
- redundancy 213
- reference bench mark 424
- reference ellipsoid
  - geocentric *see* geocentric reference ellipsoid
  - geodetic 115
- reference gravity 77
- reference sphere 117
- reference spheroid of degree  $l$  577
- refined geoidal deflection of vertical 575
- reflected wave 156
- reflection matrix 38
- reformulation of mathematical model 200
- refracted wave 156
- refraction 156
  - astronomical 302
  - horizontal 158
- index of 156
- ionospheric 160
- residual 431
- tropospheric 159
- vertical 158
- refraction correction
  - Doppler count
  - ionospheric 322
  - tropospheric 321
- levelling 431
- range 315
- residual 378
- zenith distance (Astronomical) 305
- refractive angle 158
- refractivity: specific 156
- region: confidence 222
- regional determination of gravity field parameters 544
- regression 247
  - simple 247
  - simultaneous adjustment and 258
  - two-component 247
- regular matrix 27
- regular point 41
- relative accuracy 42
- relative confidence ellipse 355
- relative confidence ellipse for displacement 642
- relative confidence ellipsoid 347
- relative confidence interval 439
- relative count 42
- relative deflection of vertical 91
- relative geoidal height 89
- relative gravity measurement 534
- relative positioning Ch. 16
  - directions to satellites 341
- horizontal
  - conformal map §16.3
  - reference ellipsoid §16.2
  - range differences 342
  - ranging to satellites 343
  - three-dimensional §16.1
    - inverse problem of 345
    - terrestrial 333
    - vertical §16.4
- relative vertical displacement 619
- relativistic effect on satellite orbit 549
- relevelled segment 612
  - observation equation for 626
  - scattered 628
- repeated gravity observation 614
- repeated point positioning 634
  - vertical 612
- representative value of observable 182
- resection 419

- residual 183  
 expected 202  
 least-squares (estimated) 207  
 standardized 229  
 statistically dependent 183  
 statistically independent 183  
 residual refraction 431  
 residual refraction correction 378  
 residual vector 188  
 resonance in ocean dynamics 128  
 resonance in satellite dynamics 560  
 resonant frequency 560  
 response of filter 251  
 response technique: local (sea level) 428  
 resurveyed network 634  
 Riemann integral 33  
 right ascension 294  
 right ascension coordinate system 293  
   mean 294  
   true 300  
 right ascension of ascending node 311  
 river discharge effect on sea level variation 425  
 robust statistics 215  
 roots of continental block 135  
 rotation: average differential 655  
 rotation matrix: fundamental 38
- sample: random 41  
 sample mean 42  
 sample point 182  
 sample variance 42  
 satellite  
   close orbiting 546  
   direction to 181  
   drag-free 548  
   high orbiting 546  
   range to 181  
 satellite altimetry 444  
 satellite distance 181  
 satellite dynamic solution for potential coefficients 562  
 satellite geometrical solution for geoid 562  
 satellite gradiometry 563  
 satellite laser ranging (SLR) 314  
 satellite orbit *see* orbit  
 satellite positioning  
   combined (with potential coefficient determination) 563  
   point 310  
     direction mode 316  
     GPS 314  
     hyperbolic mode 318  
     range difference mode 318  
     ranging 313  
 satellite laser ranging (SLR) 314  
 satellite radio ranging (SRR) 314  
 simultaneous direction and range 317  
 TRANSIT (Doppler) 318  
 relative  
   differential range mode 343  
   direction mode 341  
   GPS (differential) 343  
   range difference mode (translocation) 342  
     semidynamic 343  
     ranging 343  
     TRANSIT (Doppler) 342  
 short arc mode of 388  
 simultaneous  
   geometrical mode 385  
   kinematically constrained mode 387  
     long arc 388  
     short arc 388  
 satellite radio ranging (SRR) 314  
 satellite receiver 314  
 scalar field 36  
 scalar (inner) product 32, 35  
 scattered relevelled segments 628  
 science  
   atmospheric 22  
   computer 23  
   space 22  
 SCP *see* south celestial pole  
 sea bed effect on sea level variation 425  
 sea current effect on sea level variation 425  
 sea level  
   local instantaneous 424  
   mean 105  
     local 424  
   secular change in 148  
   tidal variation of 593  
 sea level variation 181, 148, 612  
   accelerated 632  
   annual 148  
   atmospheric pressure on 425  
   eustatic 148  
   glacial melt on 426  
   long periodic tide on 426  
   river discharge on 425  
   sea bed effect on 425  
   sea currents on 425  
   thermohaline structure on 425  
   wind stress on 425  
 sea surface topography 105  
   instantaneous 427  
 sea tide potential 601  
 second design matrix 180  
 second Love number 592

- second-order atmospheric correction to gravity  
     166  
 second-order atmospheric correction to gravity potential 166  
 second-order design problem 243  
 secondary mathematical model 181  
 secondary phase lag 421  
 sectorial contribution to tidal potential 590  
 secular change in sea level 148  
 secular orbital perturbation 557  
 secular periodic orbital variation 559  
 segment 433  
     levelling 433  
     relevelled 612  
     scattered 628  
 seismic wave 147  
 self-contained positioning system 422  
 semi-analytical stereomodel adjustment 384  
 semidynamic mode of translocation 343  
 sensor 177  
 separation of variables 36  
 sequence of elements 30  
 sequential normal equation 281  
 sequential filter 250  
 series  
     data 245  
     smoothed 249  
     eigenfunction 35  
     Fourier trigonometrical 36  
     generalized Fourier 36  
     McLaurin 33  
     random 264  
     Taylor 33  
 series of elements 30  
 series of observations in space 182  
 series of observations in time 182  
 shear rosette 653  
 shear 653  
 Shida number 591  
 short arc mode of satellite positioning 388  
 short-arc orbit prediction 554  
 short line formula 353  
 short orbital arc 387  
 short periodic orbital variation 559  
 sidereal day 59  
 sidereal time  
     Greenwich apparent 298  
     local apparent 302  
 sidereal year 58  
 signal 184  
 significance level 221  
     critical 223  
 similarity transformation 383  
 simple adjustment 258  
 simple displacement model 644  
 simple harmonic motion 35  
 simple regression 247  
 simple statistical hypothesis 220  
 simulated two-dimensional network 412  
 simultaneous adjustment and regression 258  
 simultaneous (in-context) confidence ellipse 408  
 simultaneous (in-context) confidence ellipsoid 392  
 simultaneous (in-context) confidence interval 439  
 simultaneous direction and range mathematical model 317  
 simultaneous linear equations 28  
 simultaneous probability 44  
 singular mathematical model 272  
 singular matrix 27  
 singular value analysis 208  
 singularity  
     diagnosis of 272  
     problems with 272  
     surface integral 540  
 skew-normal correction 349  
 slip displacement model 647  
 SLR *see* satellite laser ranging  
 smoothed data series 249  
 smoothing 249  
 Snell law 158  
 snow loading 133  
 solar day 59  
 solar radiation pressure 549  
 solution 177  
     best 194  
     complete 35  
     least-squares 200, 205  
     minimum constraint 273  
     non-linear 192  
     overconstrained 273  
     overdetermined 193  
     particular 35  
     underdetermined 193  
     unique 191  
 solution space 178  
 Somigliana formula 482  
 sonar device 442  
 sound 161  
 sounding 442  
 sounding datum 441  
 south celestial pole 292  
 space  
     frequency 253  
     geometrical 37  
     Hilbert 196  
     L<sub>2</sub> 196  
     linear (vector) 25  
     locally compact 30

- metric 194  
 model 178  
 normed 194  
 observation 178  
 parameter 178  
 prediction 261  
 probability 215  
 solution 178  
 vector (linear) 25
- space probe** 547  
**space science** 22  
**spatial curve** 39  
**spatial distance** 181  
**spatial prediction of vertical crustal velocity** 627  
**spatial profile** 625  
**spatially constrained displacement model** 647  
**spatially continuous displacement model** 648  
**specific refractivity** 156  
**spectral analysis** 252  
**spectrum** 254
- flat 256
  - Fourier 254
  - least-squares 256
  - line 256
  - maximum cutoff 258
  - power 258
- sphere** 40
- Brillouin 467
  - celestial 292
  - reference 117
- spherical coordinate** 38  
**spherical harmonic (function)** 469
- development into 470
  - forbidden 518
  - normalized 469
  - sectorial 586
  - tesseral 469
  - zonal 469
- spherical trigonometry** 38  
**spheroid** 106
- reference 577
- spheroidal Stokes function** 578  
**spin axis: instantaneous** 59  
**spin velocity fluctuation** 68  
**spline approximation** 247  
**spreading tectonic plate boundary** 140  
**square matrix** 26  
**square root law** 433  
**SRR** *see* satellite radio ranging  
**standard deviation** 42  
**standard deviation of velocity surface** 629  
**standard point confidence ellipse** 355  
**standard point confidence ellipsoid** 347
- standard technique for horizontal datum positioning** 396  
**standard time** 303  
**standardization of random variable** 234  
**standardized residual** 229  
**star near culmination** 308  
**star near elongation** 309  
**state vector** 184  
**statistical levelling loop misclosure** 431  
**statistical dependence** 43
- total 186
- statistical hypothesis** 220
- alternative 220
  - composite 220
  - null 220
  - simple 220
  - test of 221
- statistical independence** 43  
**statistical test**
  - chi-squared goodness of fit 227
  - in-context 229
  - most powerful 222
  - multivariate 233
  - one-tailed 224
  - out-of-context 229
  - power of 222
  - two-tailed 224
  - univariate 225
- statistical test of quadratic form of misclosures** 237  
**statistical test of quadratic form of residuals** 237  
**statistical test on mean**
  - normal 227
  - Student *t* 227
- statistical test on variance** 227  
**statistical unhomogeneity** 234  
**statistical weight** 215  
**statistically dependent residuals** 183  
**statistically independent residuals** 183  
**statistic** 220  
**statistics**
  - Bayesian 215
  - mathematical 214
  - non-parametric 215
  - parametric 214
  - robust 215
- stereomodel** 382  
**steric levellin** 427  
**stochastical model** 183  
**stochastical multivariate** 42  
**stochastical variable** 42  
**Stokes formula** 520  
**Stokes function** 519
- spheroidal 578

- Stokes integral 519  
 Stokes solution of geodetic boundary value problem §22.1  
 strain  
     horizontal  
         polar motion 608  
         tidal 596  
     network 414  
     principal axis of 653  
     surface tidal 595  
 strain conic 652  
 strain gauge 637  
 strain tensor 594  
 stratopause 152  
 stratosphere 152  
 strength of network 409  
 stress: polar motion 609  
     tidal 590  
     wind *see* wind stress  
 Student t test on mean 227  
 Sturm-Liouville boundary value problem 35  
 subset analysis 232  
 subsidence: anomalous 146  
 summation normal equation 282  
 surface 40  
     closed 34  
     coordinate 40  
     deflection difference 574  
     equipotential 84  
     isobaric 154  
     level 85  
         vertical crustal velocity 628  
 surface deflection of vertical 91, 491  
 surface density 486  
 surface density function 486  
 surface gravity anomaly 490  
 surface integral 34  
 surface strain (tidal) 595  
 surface topography: sea 105  
 surface wave (electromagnetic propagation) 422  
 surveying 19  
     control 19  
 surveying engineer 50  
 surveying technician 50  
 symmetrical deformation 649  
 symmetrical filter 251  
 symmetrical matrix 26  
 system  
     acoustic positioning 422  
     coordinate *see* coordinate system  
     Global Positioning (GPS) 314  
     self-contained positioning 422  
 system of coordinates *see* coordinate system
- system of normal equations 204  
     main 406  
 system of ordinary differential equations 35  
 system of simultaneous linear equations 28  
 systematic component of observable 183  
 systematic distortion in horizontal network 413  
 systematic error 198  
 systematic levelling network distortion 439  
 systematic noise 256
- $T-t$  (arc-to-chord) correction 362  
 table of observations 182  
 tangent vector 39  
 target correction: height of 350  
 Taylor series 33  
 technician 50  
 technique: local (sea level) response 428  
 technique for horizontal datum positioning 396  
 technique of moving averages 251  
 tectonic deformations §8.3  
 tectonic movements §8.3  
 tectonic plate boundary  
     converging 141  
     spreading 140  
     transcurrent 141  
 telluroid 117  
 temperature 152  
 temperature gradient 153  
 temporal profile 625  
 temporal variations of positions and gravity field 46  
 temporally constrained displacement model 647  
 tensor  
     metric 196  
     strain 594  
 tensor of inertia 63  
 terrain tilt: tidal variation of 598  
 terrestrial coordinate system  
     conventional 296  
     instantaneous 297  
 terrestrial relative positioning 335  
 tesseral harmonic 469  
 test of statistical hypotheses *see* statistical test  
 theorem: mean value 33  
 thermal wind 163  
 thermodynamics 162  
 thermohaline structure effect on sea level variation 425  
 third Love number (Shida) 592  
 three-dimensional method for height difference determination 373  
 three-dimensional network (geodetic) 100  
     adjustment of 375  
     merger of 393

- three-dimensional positioning  
 inverse problem of 345  
 satellite *see* satellite positioning  
 terrestrial 335
- tidal acceleration 125  
 tidal constant: Doodson 589  
 tidal deformation 124  
 tidal displacement vector 595  
 tidal force 124  
 tidal frequency 127 *see also* tidal constituent  
 tidal gravity variation 127  
 tidal horizontal strain 596  
 tidal period 127  
 tidal orbital perturbation 549  
 tidal potential 125  
 luni-solar 127  
 tidal stress 590  
 tidal tilt 127  
 tidal uplift 127  
 permanent 128  
 tidal variation of absolute gravity 593  
 tidal variation of astronomical deflection of vertical 597  
 tidal variation of dynamically undisturbed sea level 593  
 tidal variation of geodetic height 592  
 tidal variation of geoidal height 593  
 tidal variation of horizontal angle 597  
 tidal variation of horizontal distance 597  
 tidal variation of levelled height difference 598  
 tidal variation of normal height 593  
 tidal variation of observed gravity 594  
 tidal variation of observed terrain tilt 598  
 tidal variation of orthometric height 593  
 tidal variation of vertical angle 599  
 tidal water  
 gravitational attraction of 600  
 indirect effect of 600  
 loading of 132
- tide  
 atmospheric 163  
 body 590  
 water 128
- tilt  
 polar motion variation of 608  
 tidal 127  
 tidal variation of observed terrain 598  
 variation of 181
- time 181  
 atomic 303  
 ephemeris 303  
 Greenwich apparent sidereal 298  
 local apparent sidereal 302
- standard 303  
 universal 303
- time delay (in VLBI) 344  
 time variation of sea level 148
- Tissot indicatrix 41
- topocentric coordinate system 295  
 topocentric datum position parameter *see* datum (position) parameter  
 topocentric vector 317
- topographical effect on curvature of plumbline 508  
 topographical effect on observed gravity 509  
 topographical gravity gradient 513  
 topographical noise in deflection of vertical 568  
 topography: sea surface 105
- torsion 39  
 torsion balance 499  
 total derivative 31  
 total differential 31  
 total surface gravity gradient 513  
 total shear 653  
 total statistical dependence 186  
 TRA *see* true right ascension coordinate system  
 trace of matrix 27  
 transcurrent tectonic plate boundary 141  
 transformation  
 diffeomorphic 355  
 Fourier 254  
 complex 255  
 Householder orthogonal 208  
 locally linear 39  
 similarity 383  
 TRANSIT (satellite system) 318  
 transformation between (Cartesian) coordinate systems 37  
 transition function 184  
 transition matrix 184  
 transition probability function 185  
 translation component: datum 327  
 translation vector 38  
 translocation *see* satellite positioning  
 tremors 147  
 trend 183  
 analysis of 245
- triangular matrix  
 lower 27  
 upper 27
- triaxial ellipsoid 107
- trigonometrical height difference determination 364  
 trigonometrical polynomial 252  
 trigonometrical series: Fourier 36  
 trigonometrical term

- amplitude of 253
- phase of 253
- tropopause 152
- troposphere 152
- tropospheric correction to Doppler count 321
- tropospheric delay correction 315
- tropospheric refraction 159
- true (orbital) anomaly 310
- true right ascension coordinate system (TRA) 300
- truncation coefficient 543
- tsunamis 147
- two-component adjustment 259
- two-component regression 247
- two-dimensional network 412 *see also* horizontal network
- two-media photography 442
- two-sided probability value 224
- two-tailed test 224
- unbiasness with respect to  $\theta_1$  219
- unbiasness with respect to  $\theta_2$  219
- underdetermined solution 193
- unhomogeneity: statistical 234
- unhomogeneous partial differential equation 36
- uniform approximation 247
- uniform probability density function 42
- unique solution 191
- unit matrix 27
- univariate statistical test 225
- universal gravitation: law of 70
- universal time 303
- unknown parameter 177
  - least-squares estimate of 205
- uplift
  - anomalous 146
  - polar motion 607
  - tidal 127
    - permanent 128
- upper triangular matrix 27
- urban management 20
- UT *see* universal time
- value
  - boundary 35
  - functional 31
  - initial 35
- value of observable
  - estimated 188
  - exact 188
  - expected 183
  - representative 182
- Vandermonde matrix 26
- variable(s)
- dummy 33
- random 42
- scalar
  - function of 31
- separation of 36
- stochastical 42
- vector
  - function of 31
- variance
  - design 227
  - probability density function 42
  - sample 42
- variance factor 211
  - a posteriori 213
- variation *see also* correction
  - distance 181
  - earth's spin velocity 68
  - gravity *see* gravity variation
  - lake level 613
  - long periodic orbital 559
  - sea level 181, 612
    - accelerated 632
    - eustatic 148
    - temporal 147
  - secular periodic orbital 559
  - short periodic orbital 559
  - temporal (of positions and gravity field) 46
  - tidal *see* tidal variation
  - tidal gravity, 127
  - tilt 181, 614
- variation function 205
- varying earth's spin velocity induced deformation 609
- vector 25
  - binormal 39
  - constant 178
  - correction 205
  - gravity 75
  - interstation 336
  - misclosure 203
  - normal 39
  - normal gravity 78
  - parameter 188
  - position 37
  - radius 37
  - residual 188
  - state 184
  - tangent 39
  - tidal displacement 595
  - topocentric 317
  - translation 38
- vector analysis 31
- vector argument 26
- vector differential equation 35

- vector field 36  
 vector function 30  
     integral of 33  
 vector product 32  
 vector (linear) space 25  
 vector variable 31  
 velocity  
     angular 65  
     constant vertical 623  
     crustal *see* crustal movement  
     group 160  
     point 626  
     vertical 626, 627  
 Vening Meinesz formula 521  
 Vening Meinesz function 522  
 vernal point 61  
 vertical (the) 91  
 vertical: deflection of 91  
 vertical angle 294  
     geodetic 328  
     tidal variation of 599  
 vertical angle observation equation 377  
 vertical angle to star 181  
 vertical circle 293  
 vertical crustal velocity  
     constant 623  
     spatial prediction of 627  
 vertical crustal velocity surface 628  
 vertical direction 85  
 vertical displacement 615  
     absolute 619  
     four-dimensional model for 630  
     relative 619  
 vertical geodetic datum 98  
 vertical crustal movement  
     accelerated 625  
     linear 622  
 vertical positioning 612  
 vertical refraction 158  
 vertical temperature gradient 153  
 vertical crustal velocity 626, 627  
 very long base line interferometry 341  
 Vignal orthometric height 373  
 VLBI *see* very long base line interferometry  
 volume integral 34  
 water loading: tidal 132  
 water reservoir loading 132  
 water tide 128  
 water transfer 440  
 wave 253  
     direct (of electromagnetic propagation) 155  
     gravitational 147  
     ground (of electromagnetic propagation) 155  
     reflected (of electromagnetic propagation) 156  
     refracted (of electromagnetic propagation) 156  
     seismic 147  
     surface (of electromagnetic propagation) 422  
 weight: statistical 215  
 weight kernel 543  
 weight matrix 211  
 weighted constraint 270  
 weighted residual 201  
 west direction 293  
 white noise 256  
 Wiener-Kolmogorov formula 264  
 wind  
     geostrophic 162  
     thermal 163  
 wind stress effect on sea level variation 425  
 wobble (motion): polar 66  
 wobble excitation 66  
 year: sidereal 58  
 zenith distance 294  
     astronomical 294  
     geodetic 328  
     meridian 306  
 zenith distance astronomical refraction correction 305  
 zenith of observer 293  
 zonal contribution (to tidal potential) 590  
 zonal harmonic 469  
 zonal irregularity (of gravitational field) 549

## Books in Print Detailed Record

**Geodesy:**  
The Concepts

P Vanicek; E J Krakiwsky

1986-11 2nd ed., Revised

English Book 696 p.; ill

San Diego :: North Holland [Imprint]; Elsevier Science &amp; Technology Books ; ISBN: 0444877754 (Trade Cloth)

[GET THIS ITEM](#)External Resources: • [FIND RELATED](#)More Like This: [Advanced options ...](#)

Copyright: Books In Print, (c) 2005 R.R. Bowker LLC

**Title:** Geodesy:  
The Concepts**Author(s):** [Vanicek, P.](#), Author; [Krakiwsky, E. J.](#), Author**Publication:** San Diego : North Holland [Imprint]; Elsevier Science & Technology Books [Publisher Record](#)**Edition:** 2nd ed., Revised**Year:** Nov. 1986**Description:** 696 p., ill**Status:** Out of Print**Language:** English**Standard No:** **ISBN:** 0444877754 (Trade Cloth); **Other:** 9780444877758 (EAN) **LCCN:** 85-10156