Related Pergamon Titles of Interest

Books

ALLUM
Photogeology and Regional Mapping

GHOSH
Analytical Photogrammetry 2nd Edition

LISLE
Geological Structures and Maps

MALING
Measurements from Maps

ROBERTS
Introduction to Geological Maps and Structures

Journals

International Journal of Rock Mechanics and Mining Sciences

Journal of Geodynamics

Journal of Structural Geology

Coordinate Systems and Map Projections

SECOND EDITION

by

D. H. MALING
Formerly University of Wales

PERGAMON PRESS
OXFORD · NEW YORK · SEOUL · TOKYO

U.K.	Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, England
U.S.A.	Pergamon Press Inc., 395 Saw Mill River Road, Elmsford, New York 10523, U.S.A.
KOREA	Pergamon Press Korea, KPO Box 315, Seoul 110-603, Korea
JAPAN	Pergamon Press Japan, Tsunashima Building Annex, 3-20-12 Yushima, Bunkyo-ku, Tokyo 113, Japan

First edition 1973

Reprinted 1980

Second edition 1992

Library of Congress Cataloging-in-Publication Data

Maling, D. H.
Coordinate systems and map projections by D. H. Maling – 2nd ed.
p. cm.
Includes bibliographical references and index.
1. Map projection. 2. Grids (Cartography) I. Title.
GA110.M32 1991
526'.8 – dc20 91-9291

British Library Cataloguing in Publication Data

Maling, D. H. (Derek Hylton)
Coordinate systems and map projections. – 2nd ed.
1. Map projections
I. Title
526.8

ISBN 0-08-037234-1

The cover illustration is an extract from the map compiled by D. Eckhardt and published as the cover illustration for *EOS, Transactions of the American Geophysical Union*, Vol. 64, No. 25, 1983. Copyright © American Geophysical Union

Printed in Great Britain by B.P.C.C. Wheatons Ltd., Exeter

Preface

The original version of this book was written in response to a need for suitable textbooks to cover the requirements of the British national and technical certificates in cartography and surveying which had been created in the early 1970s. It was intended that the British Cartographic Society should sponsor publication of a series of titles, but, in the event, only this book ever appeared. In the Preface to the 1973 edition I described the reasons for writing this book as follows:

> It can be argued that the subject of Map Projections is better documented than some other fields in cartography: why then produce another book on the subject rather than concentrate on these other fields? There are two reasons why this is desirable. The first is that a textbook for the professional cartographer might reasonably be expected to be up to date in its treatment of the practical tasks of choosing projections for specific purposes, computing and plotting them as the preliminary to compilation. Few of the textbooks which are available satisfy these needs. The second reason is that a book of somewhat higher standard is needed for the professional cartographer of tomorrow than has hitherto been regarded as adequate for geography students. Very little has been published in Britain since the beginning of the twentieth century which treats with the mathematics of map projections at an intellectual level higher than the requirements for plane geometry and trigonometry associated with the Ordinary Level Syllabus of the General Certificate of Education. Consequently the subject of map projections often appears to the intelligent outsider as a rag-bag of separate and apparently unrelated geometrical exercises which has very little to do with the kind of map projections which are used for published maps. The weaknesses in the systems of classification evinced by many English textbooks suggest that the relationship between different projections is also not very clear to the authors. Analysis of the distortions and deformations which are inherent to all map projections are usually dealt with qualitatively rather than quantitatively. The methods of construction which are described are those of the schoolroom rather than the drawing office.

During the years which followed *Coordinate Systems and Map Projections* was much used by many people who had occasion to learn about map projections but who were not necessarily National Certificate students. For example it was used as an introductory textbook for several university geography courses. Since the feedback to the author was coming from these other sources, rather than the National Certificate courses, it became apparent that a much more comprehensive work was needed. There have been particularly insistent demands for some information about the transformations from one projection to another and from one geodetic datum to another needed by surveyors and those involved with GIS.

Seventeen years have elapsed since publication of the first edition. The intervening years have witnessed a profound change in the availability and use of digital computing, with concomitant changes in cartographic practice. There has also been a technological revolution in surveying practices which owes as much to the changes in opinion about what might be needed from a survey as to the development of new instruments and methods.

Of course digital computers have been around for more than 40 years; for example the author had first used them for work on map projections in the early 1960s, and by some standards was already late on the scene. In those days the work was all done by batch processing on what we now call mainframe computer systems. The later developments of on-line working and, particularly the availability of microcomputers immediately to hand, have revolutionised the way in which we carry out computations.

A major thesis presented in Chapter 8 concerns the construction of map projections. I argue that, provided the appropriate coordinate expressions are known, construction is purely a mechanical task, whether this be done by on-line graph plotter, or by a trainee draughtsman using spring-bow dividers and a scale to plot the master grid coordinates upon a sheet of plastic. This runs counter to the tradition of learning a unique way to construct each projection which smacks of the way trade secrets used to be handed on to apprentices. Thus we might suppose that the apprentice cartographer of the late sixteenth or seventeenth centuries would be regarded as being a right and proper person to ply his trade when he knew and could recite the rules for constructing the Stereographic, Ortelius' and the Sinusoidal projections and did not confuse one of them with another. This kind of approach differs little from that still to be found in many books on map projections.

The chapters about practical construction and computing projection coordinates are now two of the most out-of-date parts of the first edition. In the early 1970s such important innovations as the programmable pocket calculator had only just reached the UK, and the first microcomputers were still 5 years away. The methods described in the first edition still needed access to tables and, for the benefit of the majority who did not yet have access to a digital computer, there was serious consideration of the relative merits and economies to be obtained from calculating coordinates with the aid of logarithms or mechanical calculators; the comparative advantages of solving spherical triangles using haversines rather than the conventional trigonometric functions. All the coordinate computations and even the equations to determine spheroidal parameters such as radii of curvature and meridional arc distance can be done efficiently on a pocket calculator costing no more than £10. Moreover, the draughtsman may not necessarily plot the projection coordinates manually. The automated methods in cartography imagined in the 1960s

have now evolved into Geographical or Land Information Systems. Much of the work may be done using graphics packages to produce a suitable monitor display rather than a paper map. The subsequent stages of map use by comparison and evaluation of different mapped images for a particular purpose are to be found in handling GIS or LIS layers. It means that the conventional paper map is going to be replaced more and more, so that it seems possible that in another 20 years most maps as we know them will have become rarities to be consulted in libraries.

However, it is unrealistic to imagine that all GIS work can be handled by microcomputer. Since the files comprising individual layers in such systems may each comprise millions of pixels, there is need to process such data economically and in terms of transforming them geometrically, so that one layer is properly registered to another. Because of the demands upon space and storage, different and more economical numerical methods are needed to handle very large files than was traditionally used in mathematical cartography. The so-called *rubber-sheeting* methods, based upon numerical interpolation between control points, has divorced much of the work from the classic methods of computing map projections. Although some of the methods are considered towards the end of the book, the treatment is by no means exhaustive. Moreover, before plunging headlong into these methods, it is wise to heed Paul Curran's warning (Curran, 1987) that although the current geographical information systems bandwagon has much to offer by way of models and analysis:

> It has generated a plethora of empirical studies in which vast amounts of data have been sandwiched together, just because it was computationally possible to do so.

The principal growth area for new surveying practices has been at sea, where the absence of visible marks at the surface, and the need to operate out of sight of land, has led to the development of a new branch of the subject – *marine geodesy*. The impetus for this development has of course been economic; the need is for extremely accurate surveys to locate trial borings, well-heads, pipelines and drilling rigs required for the commercial exploitation of the offshore oilfields. Because some of the most valuable sites are to be found in places far beyond the conventional and practical limits of national control surveys, the need to relate such surveys to properly defined projection systems has become an important aspect of locating points or boundaries on the sea bed.

Like the first edition, the present book is concerned with principles and practical methods rather than with the formal description of the 50 or so individual map projections which have been commonly used. Thus it is not until Chapter 10 that the derivation of any specific map projection is described in any detail. Here only three are described, and primarily to demonstrate the methods of analysis which may be employed to define a map projection to meet a specific requirement. Far more important than

the facility to carry out an elaborate geometrical construction, or to treat systematically with all the important projections, is the appreciation of the patterns of distortion, and thereby to choose a suitable projection to show a particular country or distribution. Here again, the greater flexibility provided by on-line handling of GIS files gives advantages over traditional cartographic practice. In Chapter 11 the reader is warned that it is an unfamiliar luxury to choose the projection to be used as the base of a new map. This is because recompilation of detail to a different projection by traditional methods was so slow, and therefore expensive, that such a step was not undertaken lightly. Today it can be done quickly and efficiently, albeit to produce an ephemeral display upon a screen. Moreover it is now possible to consider two entirely different approaches to this problem. First, there is the time-honoured task of choosing which projection will show the desired feature with the least amount of deformation. The second is the opposite procedure; to seek to exaggerate a feature so that the resulting map is a caricature of what occurs on the ground.

One chapter which has remained virtually unaltered from the first edition concerns the use of map projections in navigation, and it contains a summary of the techniques used in Dead Reckoning navigation. Even in the early 1970s some reviewers considered it to be out of date and therefore irrelevant, but obviously missed the point that it was these traditional methods of navigation, not modern avionic systems, which made exacting demands upon chart use, and this stemmed directly from the nature of the projections used for navigation charts. Methods which did not differ greatly from those which had been used at sea in the late fifteenth century had survived from the beginning of air navigation until about 1950, and lasted for another quarter-century at sea. In the 1950s the greater speed of jet aircraft rendered graphical solutions too slow, and soon the electromagnetic version of the doppler effect was harnessed to measure track and ground speed directly. A decade later doppler was used to fix position both at sea and in the air with reference to clusters of artificial satellites, and it has now transformed geodesy and surveying, too. Since the late 1950s the character of marine transport has also changed. Nowadays there are no ocean-going passenger vessels, small coastal carriers or tramp steamers. Only huge tankers and bulk carriers remain, and these are naturally equipped with modern navigation aids. Consequently the kind of navigation carried out in a wet and pitching charthouse with a blunt pencil on a grubby chart has gone, and with it the special graphical techniques which were peculiar to the use of Mercator's projection. Yet the graphical methods of DR navigation were vitally dependent upon knowledge of the special properties of the map projections in use.

There are now few of us left who used graphical DR navigation to find

our way over mainland Europe, at night, in bad weather and against hostile opposition. Those of us who used them and survived bomber operations are all now aged about 70. When we have gone, the methods which we used will run the risk of being forgotten. Let this chapter remain unaltered as some small tribute, and a memorial for those navigators to whom a computer was a small analogue device for solving triangles of velocities, who were never really sure of their track or ground speed and to whom obtaining a fix had an entirely different meaning to its modern usage in the language.

In addition to the names of those former colleagues who helped in many ways in the production of the first edition, I would like to add that of Martin Coulson, whose advice and encouragement in recent years has been invaluable.

DEREK MALING
Defynnog, Powys
21 June 1990

The Symbols and Notation used in This Book

The number in parentheses denotes the page where the symbol was first defined or introduced.

a	major semi-axis of ellipsoid (2); maximum value for scanner angle (390); coefficient (395) (422)
a	maximum particular scale (99)
A	coefficient for coordinate transformations (38); coefficient for meridional arc distance (71); coefficient for Gauss–Krüger projection (342); A_1–A_7 Meade's coefficients for Transverse Mercator projection (444)
A	scale factor for stereographic projection (251)
b	minor semi-axis of ellipsoid (2); coefficient (282) (395) (422)
B	coefficient for coordinate transformations (38); coefficient for meridional arc distance (71); coefficient for Gauss–Krüger projection (342); B_2–B_7 Meade's coefficients for Transverse Mercator projection (446)
b	minimum particular scale (99)
c	polar radius of curvature of ellipsoid (65); constant (147); scale factor (284); coefficient (395) (422)
C	coefficient for coordinate transformations (38); coefficient for meridional arc distance (71); coefficient for Gauss–Krüger projection (342); C_1–C_5 Meade's coefficients for Transverse Mercator projection (448)
C	integration constant (199); convergence (320)
d	distance between two points on a map (283); lateral offset of scan lines (393); coefficient (395) (422); slope distance between two points on the ground (317)
d′	horizontal distance between two points (317)
d″	distance between two points corrected for height above reference figure (317)
D	coefficient for coordinate transformation (38); coefficient for meridional arc distance (71); arc distance between two points on the surface of a spheroid (76); distance from satellite to centre of earth (372); D_1–D_5

	Meade's coefficients for Transverse Mercator projection (447)
e	eccentricity of ellipsoid (64); coefficient (395)
e′	second eccentricity of ellipsoid (65)
e	scale error (109)
E	Easting coordinate (31); coefficient for coordinate transformation (38); coefficient in Sodano's formula for foot-point latitude (446)
E	Gaussian fundamental quantity of the first order (97)
f	flattening of ellipsoid (2); direction cosine (192); coefficient (395)
f	indication of a function (80)
F	coefficient for coordinate transformation (38); indication of a function (416); coefficient in Sodano's formula for foot-point latitude (446); scale factor (448); F_2–F_4 Meade's coefficients for Transverse Mercator projection (445)
F	Gaussian fundamental quantity of the first order (97)
G	coefficient in Sodano's formula for foot-point latitude (446); G_2–G_4 Meade's coefficients for Transverse Mercator projection (448)
g	direction cosine (192)
G	Gaussian fundamental quantity of first order (97)
h	direction cosine (192); height above reference surface (317)
H	height of satellite (389)
h	particular scale along the meridian through a point (98)
i	complex variable ($i^2 = -1$) (344)
J	harmonic of a satellite orbit (14)
k	particular scale along the parallel through a point (98)
k_0	particular scale along a standard parallel (204); scale factor for Transverse Mercator projection (340)
k	coefficient (390); constant (428)
K	chord distance between two points on the surface of a spheroid (75)
K	Kavraisky's constant for locating standard parallels (242)
L	$\lambda . \cos \varphi$ (443)
m	scale-factor in coordinate transformation (39); meridional arc length on spheroid (70)
m	distance between centre of map and specified particular scale (283)

n	number of points analysed or used (46); ellipsoidal parameter $(a-b)/(a+b)$ (65)
n	constant of the cone (203)
N	Northing coordinate (31)
p	parameter used in Rodrigues matrix (192)
p	area scale (104)
P	rotation and scale coefficient used in grid-on-grid transformation (42); coefficient for Gauss–Krüger projection (342); longitude function in UTM tables (362); P_0-P_5 coefficients in n and φ used for the Transverse Mercator double-projection (350)
q	parameter used in Rodrigues matrix (192); isometric latitude (216)
Q	rotation and scale coefficient used in grid-on-grid transformation (42); coefficient for Gauss–Krüger projection (342); Eastings term in UTM tables (362)
Q	meridional quadrant, being the length of the meridional arc from the equator to the geographical pole (350)
r	radius vector in polar coordinates (33); radius of a small circle (59); radius of generating globe (82); parameter used in Rodrigues matrix (192)
r	radial distance from the principal point of a photograph to an image point (373)
R	radius of a sphere (5); radius of the spherical earth (5); coefficient for Gauss–Krüger projection (342)
R	rotation matrix (42)
S	scale in the hyperbolic projection (283)
s	arc length on sphere or spheroid (23); linear distance (326)
s_e	arc length on equator (60)
s_m	arc length on meridian (59)
s_p	arc length on parallel (59)
s′	distance corresponding to s on plane (97); linear distance (322)
S	denominator of principal scale (82)
t	arc (23); maximum linear displacement (23); tan φ (345); scanning time of sensor (390); $\varepsilon^{-y/a}$ (417)
t	bearing of visual observation (327)
T	bearing of rhumb-line corresponding to t (327)
u	reduced or parametric latitude (74); coefficient used in Williams' solution of Transverse Mercator formulae (351); coefficient used in relating image points to the grid coordinates of the principal point of a photograph

(380); mathematical model comprising translation, scaling and rotation (428)

u angle on sphere measured from principal direction (101)

u' angle on plane corresponding to u and measured from principal direction (101)

v coefficient used in Williams' solution of Transverse Mercator formulae (351)

w coefficient used in Williams' solution of Transverse Mercator formulae (351)

x abscissa of cartesian coordinates (29); an angle (73); numerous combinations of symbols such as x′, x″, x_0x; X; X, etc. are defined in the text

y ordinate of cartesian coordinates (29); numerous combinations such as y; Y; Y, etc. are defined in the text

z angular distance measured at the centre of a sphere (5)

z maximum radial distance to the edge of an area to be mapped (233)

Z third dimension cartesian coordinate (74); (X+iY) (426)

Z azimuth (54)

α (alpha) grid convergence (33); angle of rotation of coordinate axes (40); bearing (54); coefficient for meridional arc distance (71); Euler's angle of rotation about the Z-axis (185); Wray's aspect parameter (190)

β (beta) coefficient for meridional arc distance (71); angle on globe between principal direction and meridian corresponding to β' on map (103); maximum angular extent of a map (110); Euler's angle of rotation about the X-axis (185); Wray's aspect parameter (190); bearing (322)

γ (gamma) angle between the axes of a plane cartesian system (43); convergence (62); coefficient for meridional arc distance (71); Euler's angle of rotation about the Y-axis (185); Wray's aspect parameter (190)

δ (delta) finite difference in the quantity which follows, e.g. $\delta\varphi$ is a difference in latitude (51); coefficient for meridional arc distance (71)

Δ definite difference in the quantity which follows, e.g. Δx is a difference in x; scale term in Rodrigues matrix (193); D/R, where D is the height of a satellite above earth's centre (372); displacements in MSS images (404)

Δt maximum difference in arc length (23)

δ minimum separation of parallel circles in an area to be mapped (233); displacement of images on aerial photographs owing to earth curvature (374)

ε (epsilon)	orbital inclination of satellite track (379); base of natural logarithms (417)
η (eta)	ordinate of curvature (333); $\eta = (\nu/\rho)^{1/2} = (e'^2 \cos^2 \varphi)^{1/2}$ (345)
θ (theta)	bearing (23); vectorial angle in polar coordinates (33); an angle (100); angle of intersection between a meridian and parallel on a map (103); angle of elevation between two ground points at different heights (317); heading of a satellite (379); scanning angle (389)
$\theta_1,\theta_2,\theta_3,\theta_4,\theta_5$	Bowring's auxiliary angles used in the determination of Gauss–Krüger equations (347)
λ (lambda)	longitude (52); independent parameter in Rodrigues matrix (192)
Λ	Wray's aspect parameter (190); longitude on an auxiliary sphere corresponding to geodetic longitude (λ) on the spheroid (350)
μ (mu)	particular scale (99); independent parameter in Rodrigues matrix (192)
μ_0	principal scale (83)
ν (nu)	transverse radius of curvature of an ellipsoid (68); independent parameter in Rodrigues matrix (192)
ξ (xi)	spherical angle used in change in aspect (192)
π (pi)	3·14159...
ρ (rho)	meridional radius of curvature of an ellipsoid (68); portion of the orbital arc of a satellite (379)
ρ	radius vector on a ground plane corresponding to r on the aerial photograph (373)
σ (sigma)	constant (282)
φ (phi)	latitude (50); geodetic latitude (66)
φ'	foot-point latitude (33); authalic latitude (415)
Φ	Wray's aspect parameter (190); latitude on an auxiliary sphere corresponding to geodetic latitude on the spheroid (350)
χ (chi)	colatitude (51)
ψ (psi)	geocentric latitude (66)
ω (omega)	maximum angular deformation (105)
Ω	Wray's aspect parameter (190)

Coordinate systems

(E, N)	grid coordinates of a point (31)
(x, y)	plane cartesian coordinates of a point (29)
(x′, y′)	master grid coordinates (182)
(X, Y, Z)	three dimensional cartesian coordinates (17); model coordinates in photogrammetry (368)

(X^*, Y^*, Z^*)	rotated three-dimensional cartesian coordinates following change in aspect (191)
(z, α)	bearing and distance (spherical polar coordinates) (178)
(r, θ)	plane polar coordinates of a point (33)
(φ, λ)	geographical coordinates (52)
(Φ, Λ)	geographical coordinates on an auxiliary sphere (350)
(r, c)	row and column coordinates locating pixels in a scanned image (394)
(u, v)	plate coordinates on an aerial photograph (380)

CHAPTER 1

The Figure of the Earth and the reference surfaces used in surveying and mapping

> The precise shape of the earth is usually referred to as a 'geoid', a term which conveys nothing beyond earth-shaped.
>
> G. P. Kellaway, *Map Projections*, 1946

Introduction

Geodesy is the science concerned with the study of the shape and size of the earth in the geometrical sense and with the study of certain physical phenomena, such as gravity, in seeking explanation of fine irregularities in the earth's shape. The subject is intimately linked with surveying and cartography. A major part of the evidence about the shape and size of the earth is based upon surveys. Indeed in some European languages the word 'geodesy' is practically equivalent to English usage of the word 'surveying'. Knowledge about the earth's size and shape is indispensable if we are to make maps of its surface. Put in the simplest form, it is necessary to know the size of the earth in order to make maps of it at known scale.

We know that the earth is a nearly spherical planet upon which are superimposed the surface irregularities created by land and sea, highland and lowland, mountains and valleys. However these topographical irregularities represent little more than a roughening of the surface. Since the radius of the earth is about 6371 km and since the major relief features do not rise more than 9 km above or fall more than 11 km below sea level, they are relatively less important than, say, the seam on a cricket ball or the indentations on the surface of a golf ball. For example, if the earth is drawn to scale as a circle of radius 6 cm, which is almost as large as the width of this page can accommodate, the variation in line thickness of the circumference which would show the entire height range from Mount Everest to the Mariana Trench at the same scale is less than 0·2 mm.

The idea that the earth is a sphere dates from the Greek geometers of

the sixth century BC. The first serious attempt to measure the size of this sphere was the classic experiment carried out by Eratosthenes in the third century BC.

Towards the end of the seventeenth century, Newton demonstrated that the concept of a truly spherical earth was inadequate to explain the equilibrium of ocean surface. He argued that because the earth is a rotating planet, the forces created by its own rotation would tend to force any liquids on the surface towards the equator. He showed, by means of a simple theoretical model, that hydrostatic equilibrium would be maintained if the equatorial axis of the earth were longer than the polar axis. This is equivalent to the statement that the body is flattened towards the poles.

The ellipsoid of rotation or spheroid

The three-dimensional body which corresponds is called an *ellipsoid of rotation*, which may be represented in section by means of an ellipse, as shown in Fig. 1.01 and elsewhere. The amount of polar flattening may be expressed by

$$f = (a-b)/a \qquad (1.01)$$

where a and b are the lengths of the major and minor semi-axes of the ellipse. The value of f, which is also known as the *ellipticity* or *compression* of the body, is always expressed as a fraction. For the earth this value is close to 1/298. We now know that the difference in length between the two semi-axes is approximately 11·5 km, or the polar axis is about 23 km shorter than the equatorial axis. It is interesting to reflect that this difference is about the same order of magnitude as the total relief variation on the earth. Thus at the approximate scale of 1/100 000 000 which represents the earth by a circle of radius 6 cm, the amount of polar flattening is also about 0·2 mm. Since 0·2 mm is also the width or gauge of line used for fine linear detail on maps, it follows that at very small scales the ellipticity of the earth is about the width of the line used to draw the elliptical section, and is therefore negligible. This is an important conclusion from the cartographic viewpoint because it permits the assumption that the earth can be regarded as truly spherical for certain purposes. We examine the validity of this assumption elsewhere (pp. 20–26). However, we must also note that any attempt to represent the terrestrial ellipsoid diagrammatically by a recognisable ellipse must involve considerable exaggeration. This, in turn, leads to possible misinterpretation of some of the illustrations depicting the geometry of the ellipsoid.

Since the ellipsoid of rotation approximates so closely to the sphere it may be called a *spheroid.* Since the flattening occurs at the poles rather than the equator, the figure may be further defined as an *oblate spheroid.*

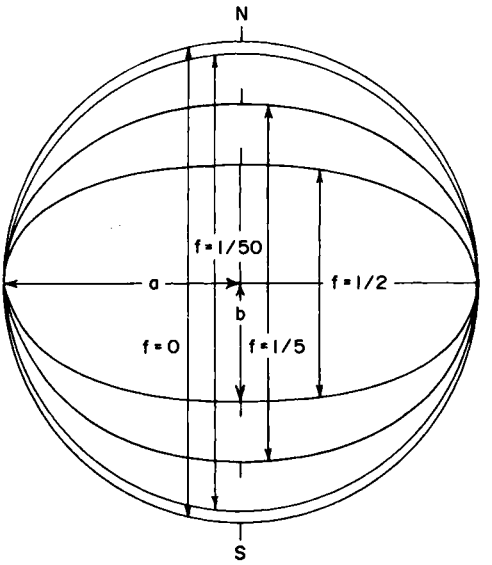


FIG. 1.01 The relationship between ellipses of different ellipticity. This diagram shows three ellipses with ellipticity f = 1/2, f = 1/5 and f = 1/50 which have the same major axis. The semi-axes of the ellipse for which f = 1/50 are a and b respectively. These ellipses are compared with a circle of radius a which is also an ellipse with ellipticity f = 0. Since most Figures of the Earth have flattening of approximately 1/298 it is clear from this figure that the terrestrial ellipsoid cannot be depicted in section at this scale in a form distinguishable from the circle. Consequently the terrestrial ellipsoid is usually represented by an ellipse with ellipticity 1/5 or thereabouts.

In the literature of surveying and cartography no real distinction can be made between the use of the two words 'ellipsoid' and 'spheroid'. Both are used indiscriminately.

Measurement of the earth's figure

Eight kinds of evidence have been used to determine the shape and size of the earth. These are:

- measurement of *astro-geodetic arcs* on the earth's surface,
- measurement of variations in gravity at the earth's surface,
- measurement of small perturbations of the moon's orbit,
- measurement of the motion of the earth's axis of rotation relative to the stars,
- measurements of the earth's gravity field from the orbits of artificial satellites,
- measurement of very long astro-geodetic arcs derived from world-wide triangulation networks,

- satellite tracking using lasers and doppler,
- measurement of the height of the sea surface using radar altimeters mounted on artificial satellites.

Certain of the methods are only of value in determining the parameter f. The purely astronomical methods, which are the third and fourth in this list, are now only of historical interest. By for the most important modern method of determination is that of radar altimetry, which has been used since 1973.

Astro-geodetic arc measurement

This is the classic method which has been used to measure both the size and shape of the earth. It is based upon comparison of the *angular distance* between two points on the earth's surface and the *linear distance* between them. The first may be determined by making astronomical observations at the two places; the second by using the precise methods of surveying referred to as *geodetic* or *first-order survey*. The *radii of curvature* of the earth may be determined from these data and finally the lengths of the semi-axes of the ellipsoid can be calculated.

If the earth were a true sphere its radius would be easily calculated, for it is a fundamental property of a sphere that all points on the surface are equidistant from its centre, i.e. it has constant radius. This is why it is possible to illustrate any section passing through the centre of a sphere by means of a circle as in Fig. 1.02. If there are two points, A and B, on the surface of the sphere with centre O, the angular distance between the points is the angle AOB measured at the centre and the arc distance

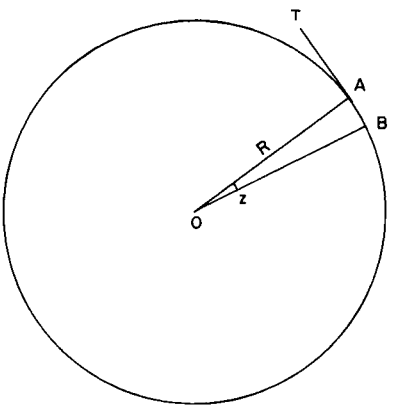


FIG. 1.02 A sphere in section, illustrating the relationship between angular distance and arc distance for all parts of the surface. AT represents a tangent to the circumference at A.

between them is the shorter part of the circumference passing through the points. The relationship between these two measurements can be determined from

$$\text{arc length } AB = \mathrm{R} \,.\, \mathrm{z} \tag{1.02}$$

where R is the radius of the sphere and z is the angle AOB expressed in radians. For example, if $\mathrm{z} = 10^\circ = 0{\cdot}174\ 53$ radians and $\mathrm{R} = 6371$ km, the arc distance $AB = 1111{\cdot}9$ km. This is constant for all values of $\mathrm{z} = 10^\circ$ on this sphere irrespective of where the arc is situated. The converse argument is used to derive the radius from the measured length of the arc and an angular measurement. Thus, if astronomical observations made at both A and B showed that they lie 10° apart and survey has established that the distance between them on the surface is 1111·9 km from equation (1.02)

$$\mathrm{R} = 1111{\cdot}9/0{\cdot}174\,53$$

$$= 6371 \text{ km}$$

The radius of the sphere has been defined as the line OA. A further property of the sphere, which may be proved from the elementary plane geometry of the circle, is that when a tangent meets a circle at the point A, the *normal* or perpendicular to that tangent passes through the centre of the circle. Thus on the sphere, OA is perpendicular to any tangent at A and if a series of tangents are drawn through A in any other directions than the section illustrated, these all lie in the same *tangent plane*.

This is important in defining the radii of curvature of an ellipsoid which are *lines perpendicular to the tangent plane at any point on the curved surface*. They are *not* represented by straight lines joining points on the surface to the geometrical centre of the body. Thus at some point A on the surface of the ellipsoid, we may imagine the tangent plane. In Fig. 1.03 the normal to this tangent plane is $AQ'Q$. A further difficulty in defining the geometry of the ellipsoid is that two separate radii may be distinguished. One of these is the radius of the arc NAE; the other is the radius of the arc which is perpendicular to NAE at A. The radii are represented in Fig. 1.03 by the lines AQ' and AQ respectively. Thus both arcs occupy the same position in space but have different lengths. Moreover the line $AQ'Q$ does not pass through the geometrical centre of the ellipse, O, except where the normal to the surface forms either NO or EON, which are the semi-axes of the figure. It follows that the radii of an ellipsoid are variable quantities. Two separate radii may be defined for each point on the surface and both of these vary with position of the point. It follows, therefore, that the linear distance corresponding to a given angular distance varies with latitude. For example, the angle $\mathrm{z} = 10^\circ$ between the points A and B near the equator represents an arc distance

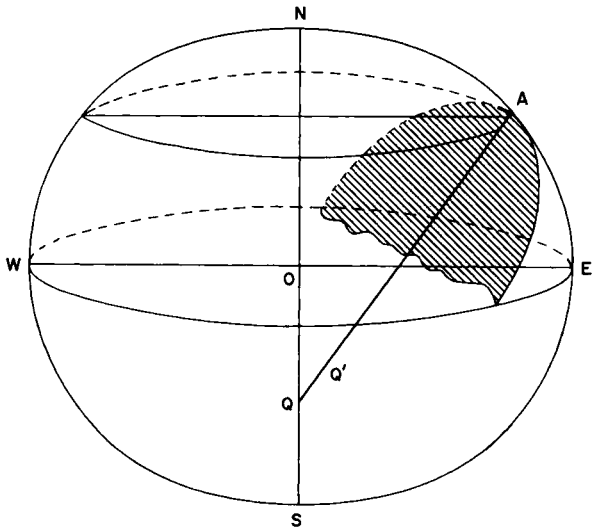


FIG. 1.03 An ellipsoid in section illustrating the meridional radius of curvature AQ' and the transverse radius of curvature AQ of the point A. The shaded plane is perpendicular to the meridian NAE through A.

of approximately 1105·6 km on the surface of the terrestrial ellipsoid, whereas the same angle between the points A' and B' near the poles corresponds to about 1169·9 km. In other words *the arc distance corresponding to a given angle increases polewards.* This relationship is shown on Fig. 1.04, but care must be taken in the interpretation of the diagram. The ellipse is shown with exaggerated compression and the directions of the radii of curvature are shown as the normals to the tangents at the four points. These must be produced to give the points of intersection at K and M' to show that $AKB = 10° = A'M'B'$. The reader should avoid making the implied comparison with Fig. 1.02, which suggests that the radii of the ellipse are the lines AK, BK etc., and hence the fallacious interpretation of them as being much greater or less than OA or OB in Fig. 1.02.

This preliminary excursion into the geometrical properties of the sphere and ellipsoid, which are examined in greater detail in Chapter 3, has been made to indicate the kind of evidence to be obtained from astro-geodetic arc measurement. The variation in arc length with latitude was one of the first important pieces of evidence to be obtained which supported Newton's theoretical gravitational model. It was obtained from the measurement of two arcs, in Peru and Lapland, by the French during the early part of the eighteenth century.

The period of greatest activity in this field of geodesy occurred during the nineteenth and early twentieth centuries. Figure 1.05 illustrates those

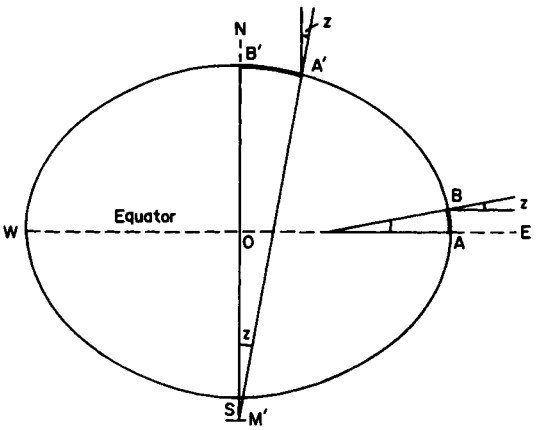


FIG. 1.04 An ellipsoid in section, illustrating how a given angular distance, z, is represented by a longer arc distance $A'B'$, near the poles than is the arc distance, AB, near the equator. Note that the radii of curvature KA and $M'A'$ also increase towards the poles but are exaggerated here owing to the exaggeration of the ellipticity of the ellipse.

arcs measured and used for the determination of different Figures of the Earth before 1914. At that time the only satisfactory method for control surveys of the requisite order of precision was by means of triangulation. The preferred type of measurement was the *arc of the meridian*, i.e. a survey made between points which differed greatly in latitude but little in longitude, so that the network of connecting triangles was aligned along the same meridian.

Note how early in the history of science some of these determinations were made. For example, the Great Trigonometrical Survey of India had measured the arc following the meridian 78°E, which crosses the centre of the subcontinent from Cape Cormorin to Kalianpur by 1825 and reached the Himalaya by 1841. Everest made the first determination of the Figure of the Earth which bears his name, and which is still in use. during a prolonged spell of sick leave from his post as Superintendent of the Great Trigonometrical Survey. For a biographical commentary on Everest and this work see Heaney (1967).

The small differences in the size and ellipticity which are shown in Table 1.01 result from subtle and small variations in the earth's figure causing it to depart from the perfect spheroid. Consequently the parameters for each Figure of the Earth depend upon which astro-geodetic arc measurements are used in the determination, and therefore the different figures tend to fit certain parts of the world better than others.

We may also observe that many of the recent figures differ by only a few centimetres in the length of the semi-major axis, a, and the flattening by 1 part in 10^{-7}. The differences are so small that it might be argued

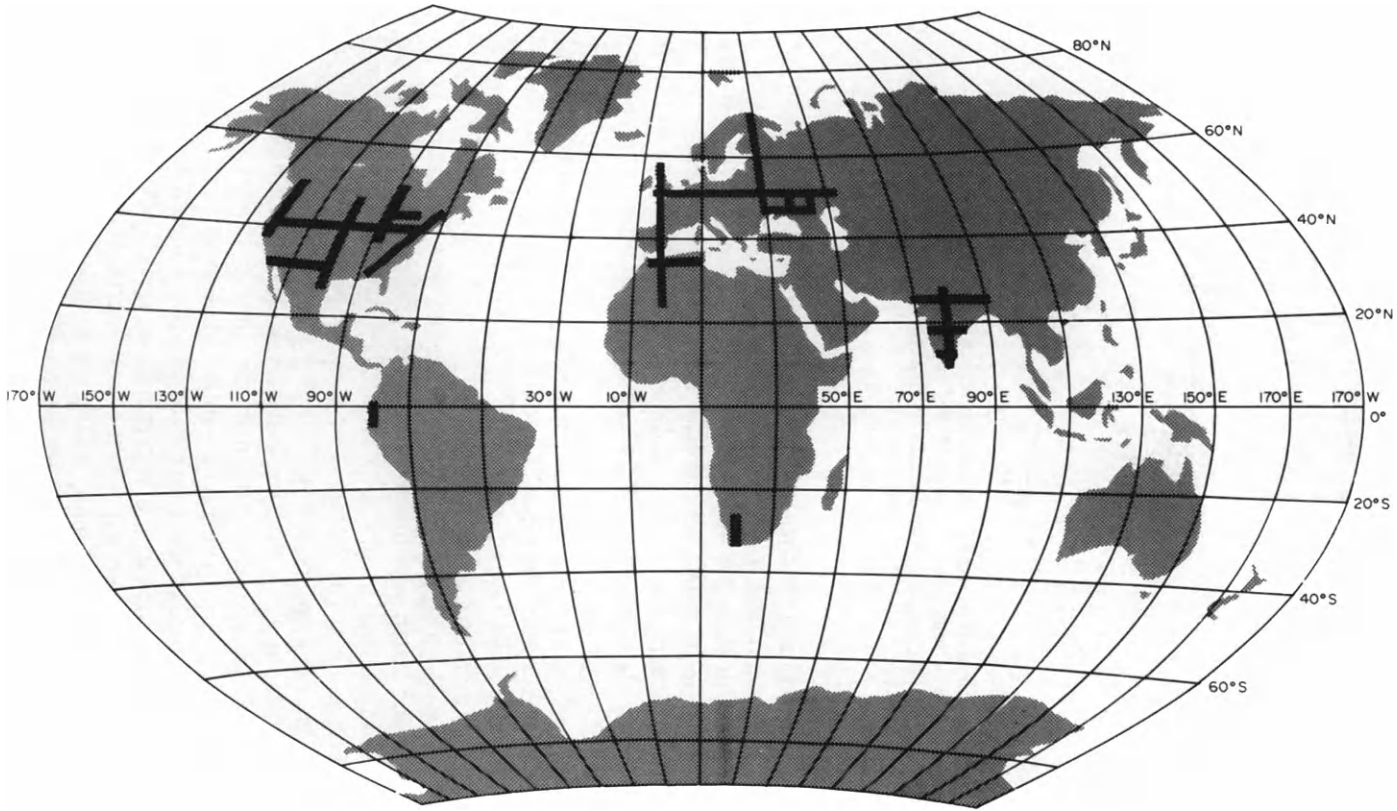


FIG. 1.05 World map showing the location of the main astro-geodetic arcs used for determination of the figure of the earth during the classical period of arc measurement before 1914. The map is based upon the Aitoff–Wagner projection (No. 39a in Appendix I) which is a member of the polyconic group of projections with equidistant spacing of the parallels along each meridian. The geographical poles are represented by curves.

whether there is any justification for regarding these as deserving separate recognition. The figures have been determined by modern methods of using tracking of artificial satellites and by direct measurement of the height of the sea surface using radar altimetry. Associated with these developments also came the methods of position fixing by measuring doppler frequency changes and therefore distances between craft and groups of artificial satellites. The systems known as the Navy Navigation System (Transit or Navstar) and the Global Positioning System (GPS) both require the motions of such a satellite (or its *ephemeris*) to be referred to a specific Figure of the Earth. Conversely, the correct figure must be used with a particular navigation system to achieve the expected accuracy.

Gravity measurements

Newton arrived at the conclusion that the earth was an ellipsoid from the theoretical consideration of the forces created by the earth's mass and rotation (see page 64). Consequently the second important line of evidence concerning the shape of the earth has been from the study of variation in gravity.

In the absolute sense gravity varies with latitude, and it was early recognised that pendulum clocks which kept good time in Europe tended to lose time near the equator.

Gravity also affects the observations made during astro-geodetic arc measurement. It is this relative aspect of gravity which is particularly important in geodesy. In order to make observations in survey and astronomy it is necessary to align the instruments to a common datum. This datum is provided by the tangent plane to the earth's curved surface at the point of observation. This plane is geometrically important and also has a physical significance because it is defined by the spirit bubble mounted on a theodolite which is adjusted by means of its footscrews until the bubble is stationary in the centre of its run. The normal to this tangent plane is defined by the plumb-line which is used to set the instrument precisely over the point from which the observations are to be made. In short, we use gravity to determine both the horizontal plane of reference and the direction of the vertical. These adjustments are normal survey practice and are especially important in geodetic measurements. Supposedly horizontal angles observed by a theodolite which is not level contain errors which consequently deform the shapes of the triangles which have been observed. This, in turn, leads to errors in the computed distances between points and therefore to error in the computed positions of the stations. Precise determination of the horizontal plane of reference is an even more vital requirement in field astronomy because position is determined from measurements of vertical angles (or the *altitudes*) of stars. The datum for these measurements is the horizontal

TABLE 1.01 *Some of the principal determinations of the Figure of the Earth*

Date	Name	Lengths of semi-axes (m)		Ellipticity (f)	Uses
		Major (a)	Minor (b)		
1687	Newton	—	—	1/231	—
1738	Maupertuis	6 397 300		1/191	—
1749	Bouguer	—	—	1/266	—
1799	Commission des Poids et Mésures	6 375 738·7		1/334·29	—
1810	Delambre	6 376 428		1/311·5	Belgium†
1817	Plessis	6 376 523	6 355 863	1/308·65	France†
1830	**Everest**	**6 377 276·345**	**6 356 075·413**	**1/300·8017**	**India, Burma, Ceylon, Malaysia (part)**
1841	**Bessel**	**6 377 397·155**	**6 356 078·963**	**1/299·1528**	**Most parts of Central Europe, Chile, China, Southeast Asia**
1849	**Airy**	**6 377 563·396**	**6 356 256·9**	**1/299·3249**	**Great Britain**
1858	Clarke	6 378 294	6 356 618	1/294·261	Australia†
1866	**Clarke**	**6 378 206·4**		**1/298·97866982**	**North America (NAD 27)**
1876	Andrae	6 377 104·43	6 356 762	1/300·0	Denmark, Iceland†
1880	**Clarke**	**6 378 249·17**	**6 356 514·9**	**1/293·465**	**France, most of the African continent**
1907	Helmert	6 378 200	6 356 818	1/298·30	Egypt†
1909	**Hayford = International (1924)**	**6 378 388**	**6 356 911·946**	**1/297·0**	**Whole world excluding North America, Africa and a few other small areas**

1940	**Krasovsky**	**6 378 245**	**6 356 863·019**	**1/298·3**	**USSR and all other communist countries**
1956	Army Map Service	6 378 270·0	6 356 794·343	1/297·0	
1960	Fischer (Mercury Datum)	6 378 166·0	6 356 784·284	1/298·3	
1960	**World Geodetic System (WGS60)**	**6 378 165**	**6 356 783**	**1/298·3**	**Australia†**
1961	Kaula	6 378 163		1/298·24	
1965	**International Astronomical Union (IAU65)**	**6 378 160**	**6 356 774·719**	**1/298·25**	**Australia**
1965	Applied Physics Laboratory (APL 4·5)	6 378 137		1/298·25	
1965	Naval Weapons Laboratory (NWL9D)	6 378 145	6 356 759·8	1/298·25	
1966	**World Geodetic System (WGS66)**	**6 378 145**	**6 356 760**	**1/298·25**	
1967	**Geodetic Reference System (GRS67) or International Astronomical Union (IAU68)**	**6 378 160·00**	**6 356 774·161**	**1/298·2472**	
1968	Fischer (Modified Mercury Datum)	6 378 150	6 356 768·955	1/298·3	—
1972	World Geodetic System (WGS72)	6 378 135	6 356 750·52	1/298·26	—
1979	Lerch *et al.*	6 378 139		1/298·257	
1980	**Geodetic Reference System (GRS80) or New International**	**6 378 137·00**	**6 356 752·3141**	**1/298·257222101**	**North America (NAD 83)**
1984	World Geodetic System	6 378 137·00		1/298·257223563	
1985	Engelis	6 378 136·05		1/298·2566	

The principal figures used for topographic mapping are tabulated in **Bold type**.

Obsolete figures formerly used for some mapping are denoted †.

There are numerous inconsistencies in use for different map series by certain national survey organisations. In military mapping other doctrines prevail. For example NATO policy for use with the UTM system (see pp. 357–360) favours use of the International Spheroid almost everywhere except North America, Africa and parts of Asia. Soviet policy has been to adopt the Krasovsky figure for all Warsaw Pact mapping.

plane indicated by the spirit bubble, or an artificial horizon formed by a liquid such as a dish of mercury which takes a horizontal position through gravitational attraction. The consequence of a slight inclination of either plane of reference leads to incorrect measurement of the vertical angle, and therefore to the determination of an incorrect astronomical position for the instrument.

The geoid

If the height of each observation station is reduced to sea-level, then by virtue of the fact that the instruments have always been carefully levelled, this is equivalent to stating that the observations have all been reduced to the same *equipotential surface* where the spirit bubble is always at rest. This surface is known as the geoid. It can be likened to the surface of an imaginary world ocean without land, waves, swell, tides or currents.

If the earth were such a homogeneous body, then from classical gravitational theory the surface of the geoid would coincide everywhere with the surface of an ellipsoid of rotation. However, this is not so. The geological history of the earth has led to irregular distribution of crustal rocks having different densities. The denser rocks exert their own attraction upon a spirit bubble, although this is small compared with the main gravitational component. Thus an instrument may *appear* to be level because the spirit bubble is at rest in the centre of its run, but the plumb-line is not normal to the spheroid for it is deflected slightly towards the areas of greater rock density. Since the amount of deflection varies from place to place it follows that the geoid has an undulating surface. Figure 1.06 illustrates how these undulations occur. Since all the observations have been made with reference to the geoid, additional measurements of the *gravity anomalies* which are present can be used to correct for and increase knowledge about the location of the undulations of its surface. Stokes first demonstrated these principles in 1849 and methods of correcting for anomalies have been used since 1855, when Pratt attempted to account for discrepancies in the position of Kalianpur observed in the astro-geodetic arc measured by the Great Trigonometrical Survey of India. The attempts to explain this and similar inconsistencies in other arc measurements led to the formulation of the different theories of *isostasy*, which have been a major preoccupation of geodesists and also revolutionised early theories about structural geology.

It follows that the increasing refinement of determination of the Figure of the Earth, characterised by the small variation in f, obtained after 1900, is largely owing to the increasing availability of gravity data and the methods of employing these to adjust the astro-geodetic observations. By the late 1940s sufficient information about gravity anomalies had been collected to attempt the compilation of maps showing the undulations of

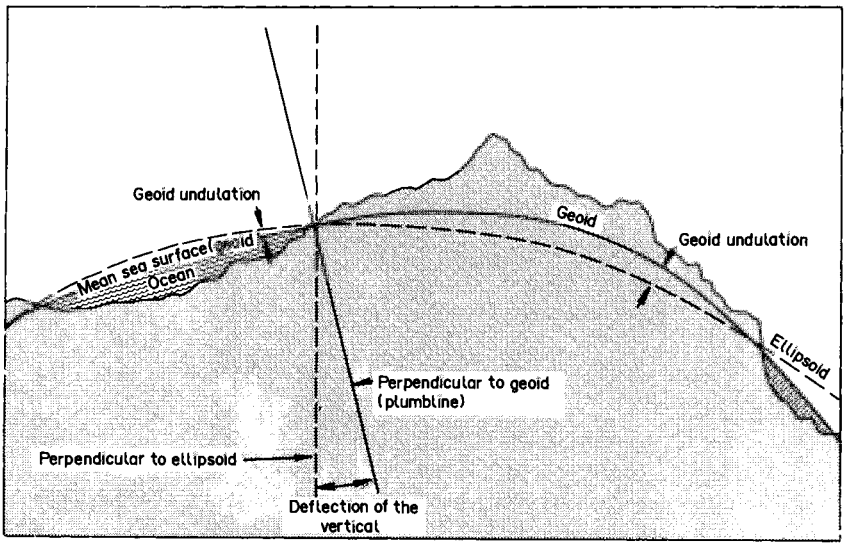


FIG. 1.06 The relationship between the geoid and reference spheroid, indicating the deflection of the perpendicular to the geoid and resulting undulations in the surface of the geoid.

the geoid by means of contours. However, these maps were confined to showing quite small parts of the northern hemisphere (USA and western Europe) where there was sufficient density of information to attempt an interpretation. Ultimately it might have been possible to proceed with such work on a world-wide basis (indeed this has been in progress ever since those days), but it would have been a very long job. At that time there was still very little information about gravity anomalies in the southern hemisphere and, moreover, there were still practical difficulties about obtaining satisfactory gravity measurements at sea. This means that there were no data from more than 70% of the earth's surface. The first successful measurements of gravity from a surface ship were only made in November 1957.

The contribution of satellite geodesy

The first artificial satellite had been launched a month earlier. This heralded a major step forward in advancement of knowledge about the earth's true shape and size, and moreover removed the dependence upon the slow acquisition of terrestrial measurements. The principal reason for this advance was that artificial satellites overcame a fundamental difficulty in deciphering the earth's gravity field, namely that because terrestrial measurements were confined within it, this made it impossible to make any external measurement of the forces. This difficulty had long been

realised; indeed, attempts had been made to employ the moon, as our natural satellite. However, the attempted measurements were somewhat insensitive because of the distance between the earth and the moon.

The evidence of satellite tracking

The earliest, and some of the most significant, information was obtained within a year or two, simply from observation of the changing orbits of the early Sputnik and Vanguard satellites. Satellite tracking has yielded much information about the gravity potential of the earth, and led eventually to remarkably detailed mapping of the geoid throughout the world (Figs 1.10 and 1.11). The second use of satellites has been to provide survey beacons which have been located high enough above the earth's surface to be simultaneously visible from places which are hundreds, or even thousands, of miles apart. Consequently these may be used to create unified and world-wide networks of geodetic stations (Fig. 1.07). This made it possible to compare astro-geodetic arcs for much greater distances on the earth's surface than had ever been accomplished in classic geodesy.

If the earth were spherical, and of homogeneous density, the orbit of a satellite would be an ellipse fixed in shape and size, and with its plane in a fixed direction in space. Any departure of the earth from a spherical form causes changes in the gravitational forces acting upon the satellite, and therefore upon its orbit. The main effect of the earth's ellipticity upon a satellite orbit is to make the plane of the orbit rotate about the earth's axis in the direction opposite to the satellite's motion, while leaving the inclination of the orbit to the equator virtually constant. This phenomenon is known as the *precession of the nodes* (Fig. 1.08). The rate of procession can be measured with extraordinarily high precision using quite simple equipment because the movement is regular and therefore it can be allowed to accumulate over long periods and therefore many orbits between observations. The value of ellipticity, obtained only a year or so after the first artificial satellite had been launched, was $f = 1/298{\cdot}24$, or practically the same as that determined by Helmert in 1907 and Krasovsky in 1940.

Study of the variations in gravity potential with latitude has led to the evaluation of a series of numerical coefficients, called J-harmonics, which describe a sequence of increasingly elaborate geometrical figures. The J_2 coefficient, which defines the ellipticity of the spheroid, is by far the most important of these, but some of the other coefficients are not wholly insignificant. They indicate that the earth is somewhat asymmetrical in section, for the North Pole lies about 10 m further from the equator than can be accounted for by ellipticity of 1/298·24, but the South Pole lies about 30 m nearer the equator than this amount of compression suggests. The resulting meridional section (Fig. 1.09) has been likened to the shape

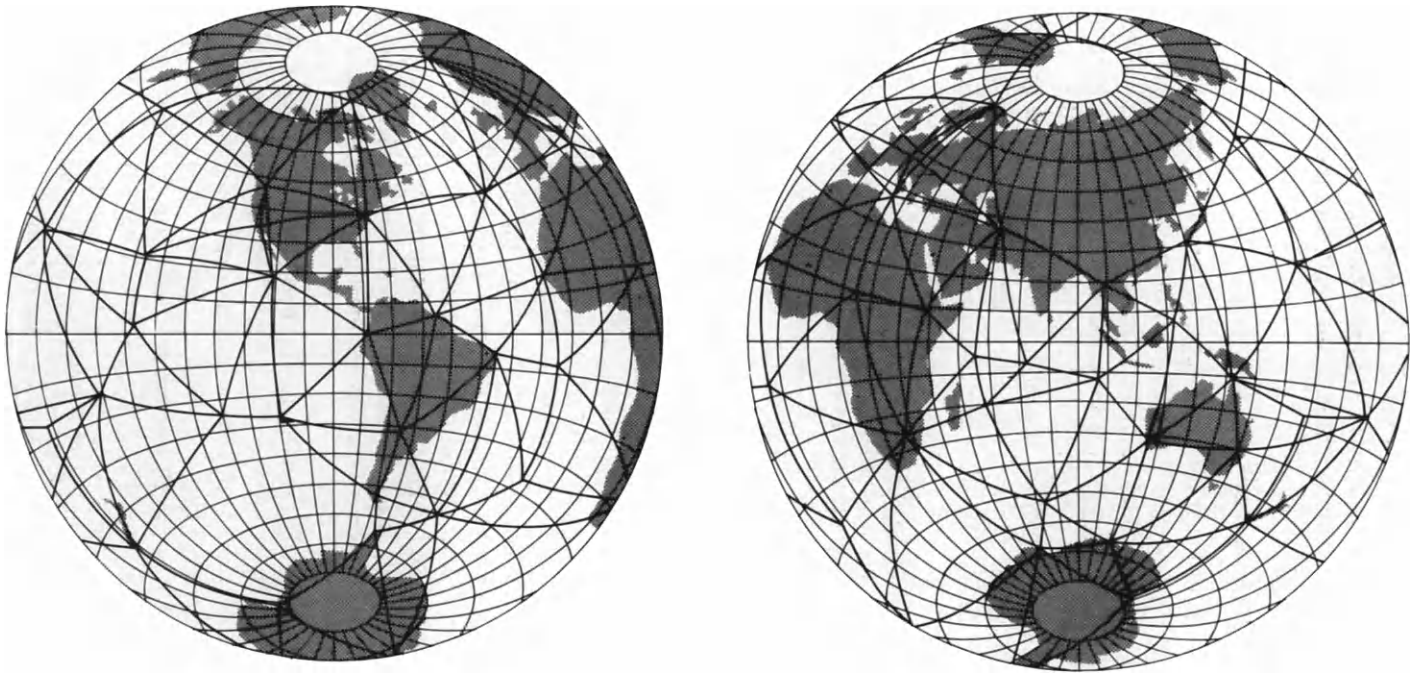


FIG. 1.07 Maps showing the world-wide BC-4 triangulation network which was observed in the early 1970s with the aid of the PAGEOS artificial satellite and ballistic camera photography as part of the US National Geodetic Satellite Program. The maps used are equatorial (transverse) aspect stereographic projections (No. 6 in Appendix I). This is a member of the azimuthal class of projections. The two maps have been extended to represent more than hemispheres ($z = 110°$) so that the edges overlap. The stereographic projection is conformal; therefore the angles within the triangulation network are correctly depicted at each observation station. Moreover all great circle arcs are shown as circular arcs on the stereographic projection. Therefore the sides of the triangulation network are represented by circular arcs.

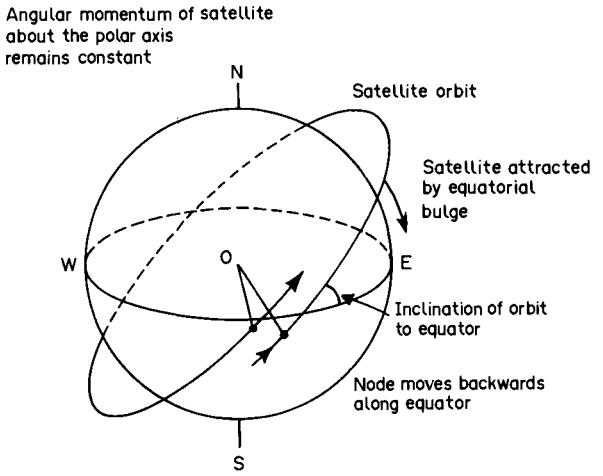


FIG. 1.08 Diagrammatic representation of the precession of the nodes. The equator-wards force, resulting from the earth's equatorial bulge, causes an artificial satellite to cross the equator on a different meridian at each successive orbit.

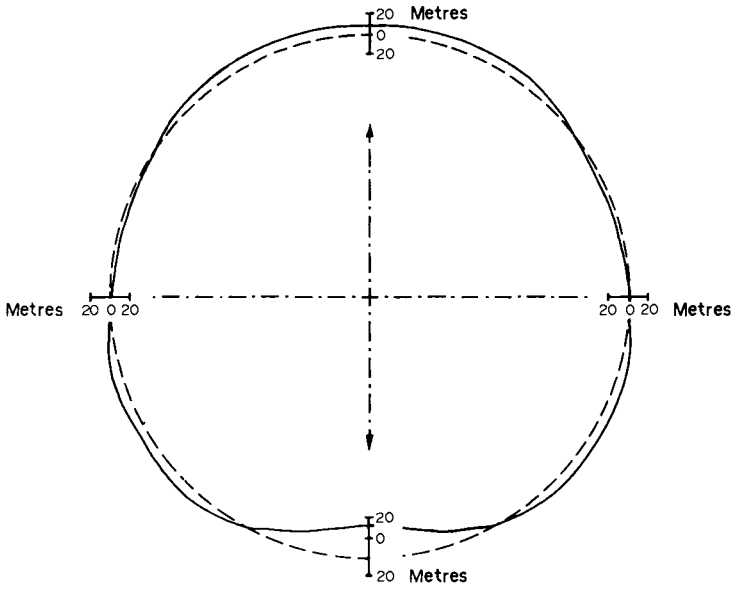
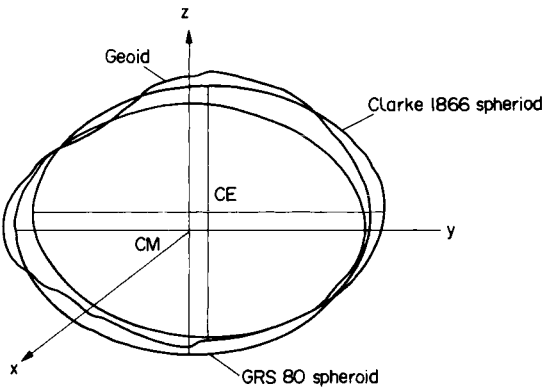


FIG. 1.09 Inferred meridional section of the earth based upon the calculation of variations in gravity potential with latitude but excluding any variation with longitude. The diagram indicates the departure (in metres) of this section (full line) from an ellipse with compression 1/298·24 (broken line).

of a pear. However, despite much publicity of this conclusion in the early days of satellite tracking, not too much importance should be placed on it, for the pear shape is an average value of the undulations of the geoid determined with reference to latitude and ignoring any variations in longitude. More important to modern concepts of geodesy were the attempts to produce a contour map of the height of the geoid for the whole world. A study first undertaken by Kaula in 1961 produced the world map illustrated in the first edition of this book. Events have proceeded so quickly that much more detailed geoid contours are now available, as illustrated in Fig. 1.11.

It is also important to appreciate that in classical geodesy the arc measurements were self-contained and isolated from one another by whole continents and oceans. Consequently the results of these arc measurements were fitted to a comparatively small portion of the spheroid, and it was impossible to relate the results precisely to the axis of rotation and the true centre of the earth. Thus a particular Figure of the Earth would not be referred to the true axis of rotation but to a parallel axis which was displaced from the true axis by a small but unknown amount, as illustrated in Fig. 1.10. For the creation of reliable satellite navigation systems the ephemeris of each satellite has to be known more precisely. This includes knowledge about the true position of the earth's centre. Consequently there has been a revolution in the concept of how the earth's figure should be defined, and a variety of new figures have emerged from these data. Modern determinations of the



CE is the centre of the ellipsoid at the intersection of the axes.
CM is the centre of mass at the origin of the XYZ coordinates.

FIG. 1.10 A comparison between the earth's figure based upon an equipotential ellipsoid, having a geocentric origin to the X,Y,Z, cartesian coordinate system, and a figure derived from classical methods of geodesy in which the centre is offset from its true position.

earth's figure from the time of GRS67 onwards are truly geocentric and based upon the theory of an *equipotential ellipsoid.* Consequently the modern trend is to describe new figures initially in geophysical terms and only later derive the various parameters to which we are accustomed. See, for example, the detailed description of the IUGG specification for GRS80 by Moritz (1980a).

Global triangulation schemes

A vital stage in satellite geodesy was therefore the accomplishment of various world-wide control surveys. The period of greatest activity in this field was in the late 1960s, during which time the whole task of providing a world-wide geodetic control network was accomplished. A variety of different techniques were employed by the different branches of the US administration involved in this renaissance of geodesy. One system favoured the use of large satellites, like the PAGEOS satellite which was a balloon that became inflated when in orbit, and therefore large enough to be simultaneously photographed against the background of stars by several BC-4 ballistic cameras. Because of the designation of the camera this project is now commonly referred to as the *BC-4 Triangulation.* The ANNA satellite contained a brilliant flashing light bright enough to be identifiable as a beacon in space. The third idea was to use electronic distance measurement to track a comparatively small reflecting satellite. This was exemplified by the SECOR system used to establish an equatorial control network round the world. Later still came the application of even more sophisticated methods of distance measurement using lasers and doppler, resulting in much greater accuracy in the methods of satellite tracking. Indeed the roles were reversed; for the positions of many modern satellites are now determined so accurately that distance measurements from clusters of them are now used to locate positions on the earth's surface. This has been developed through the various satellite navigation systems to the Global Positioning System (GPS) which promises to offer the world-wide ability to fix position with an accuracy equivalent to conventional geodetic surveys.

Satellite altimetry

Satellite altimeters directly measure the distance between a satellite and the instantaneous sea surface. By accurately determining the satellite orbit with respect to positions on the earth's surface it is possible to estimate the height of the sea surface above the reference ellipsoid. Therefore the construction of contours for the surface of the geoid can be used to estimate the deflection of the vertical at sea. The first experiments in radar altimetry in space were made from SKYLAB, launched in

November 1973. Two later satellites have so far been equipped with radar altimeters, first GEOS-3 and secondly SEASAT.

We have already likened the equipotential surface of the geoid to that of an imaginary planetary ocean. The question which naturally arises is whether the actual surface of the ocean is anything like the idealised surface, and what corrections can be applied to the natural disturbances caused by ocean currents, tides and other surface displacements in order to describe the geoid.

The GEOS-3 mission was designed to improve knowledge of the earth's gravitational field, the size and shape of the terrestrial geoid, deep ocean tides, sea state, current structure, crustal structure, solid earth dynamics, and remote sensing technology. The GEOS-3 altimeter was designed to provide the means for establishing the feasibility for directly measuring some of these variables. In every respect the altimeter far exceeded its expectations. For example, although the system was designed for a 1-year lifetime, the satellite was still operational after more than $3\frac{1}{2}$ years in orbit. In addition, the altimeter showed that it was capable of providing valid measurements over land and ice. Neither of these capabilities had been predicted prior to launch.

The second reason for the success of altimetric measurements is the speed with which the information may be collected by satellite compared with conventional marine gravity measurements. A research ship on a cruise to the Antarctic might be away for 6 months, but only a small proportion of that time will be spent making observations in the intended working area. By contrast a satellite will not only make the journey 14 times in one day, storing its results and transmitting them to a convenient ground station, but will also sense all the other oceans several times in the same day.

The principal limitation in the use of GEOS-3 altimetry was the restricted cover of the world's oceans which could be sampled. These data were largely confined to the North Atlantic, the Gulf of Mexico, North Pacific and the Bering Sea. The restricted cover was owing to the small number and the location of ground stations capable of receiving signals from the satellite.

An important method of analysis of the altimeter records is the study of those crossover points where the height of a point on the sea surface has been measured when the satellite has occupied different orbits (Marsh *et al.*, 1982a,b) allowing the precision of the surfaces to approach that of the measurements themselves (25 cm for GEOS-3, better than 10 cm for SEASAT). Analysis of the sea height residuals at the crossing points of the satellite arcs provides information about the long-term variability of sea height in these regions.

A more sensitive radar altimeter was fitted aboard SEASAT, which was launched on 27 June 1978; the network of receiving stations had also

been much extended. The satellite operated successfully until 10 October 1978, when a power failure brought transmission to a stop. A mission overview has been given by Lame and Born (1982), who have shown that, despite its short lifetime, SEASAT acquired a wealth of data on sea-surface winds and temperature, ocean wave heights, internal waves, atmospheric water content, sea ice, topography of the ocean surface and shape of the marine geoid. Analysis of the output from the radar altimeter was one of the most important aspects because most of the world's oceans were sampled; therefore better estimations were obtained for the slope of the marine geoid for the world as a whole.

Concurrent with these developments, various attempts were made to produce increasingly more sophisticated models of the geoid. These have been conventionally named after the American laboratories which have undertaken the study; notably the Smithsonian model earths, labelled SAO, after the Smithsonian Astrophysical Observatory and the GEM, or Goddard Earth Models, after the Goddard Space Center operated by NASA.

The choice of a suitable reference surface for mapping

Because we now know that the geoid is a complicated body, we must enquire how it should be described mathematically for the practical purposes of mapping. Since there is no merit to be gained from increasing the mathematical complexity of a solution beyond defining those irregularities which have practical significance, it is desirable to consider the possibility of using various *reference surfaces* which describe the shape and size of the earth adequately for different purposes. The variations illustrated by the contour pattern in Fig. 1.11 may amount to only a few metres but they are of considerable importance to the study of dynamic geodesy and some branches of geophysics. For work in these fields there are cogent reasons for defining as a reference surface a *triaxial ellipsoid* in which the observed undulations along the equator may also be fitted to an ellipse. However, these variations in the geoid are practically negligible for most other kinds of survey and in cartography.

Thus we may simplify the problem and consider three different ways in which we may define the shape and size of the earth for different purposes in surveying and mapping. These are:

1. a plane which is tangential to the earth at some point;
2. a perfect sphere of suitable radius;
3. an ellipsoid of rotation of suitable dimensions and ellipticity.

They are listed in ascending order of refinement. Thus a suitable ellipsoid fits the shape of the geoid better than does a perfect sphere of equivalent size. The sphere, in turn, is a better approximation of the curved surface

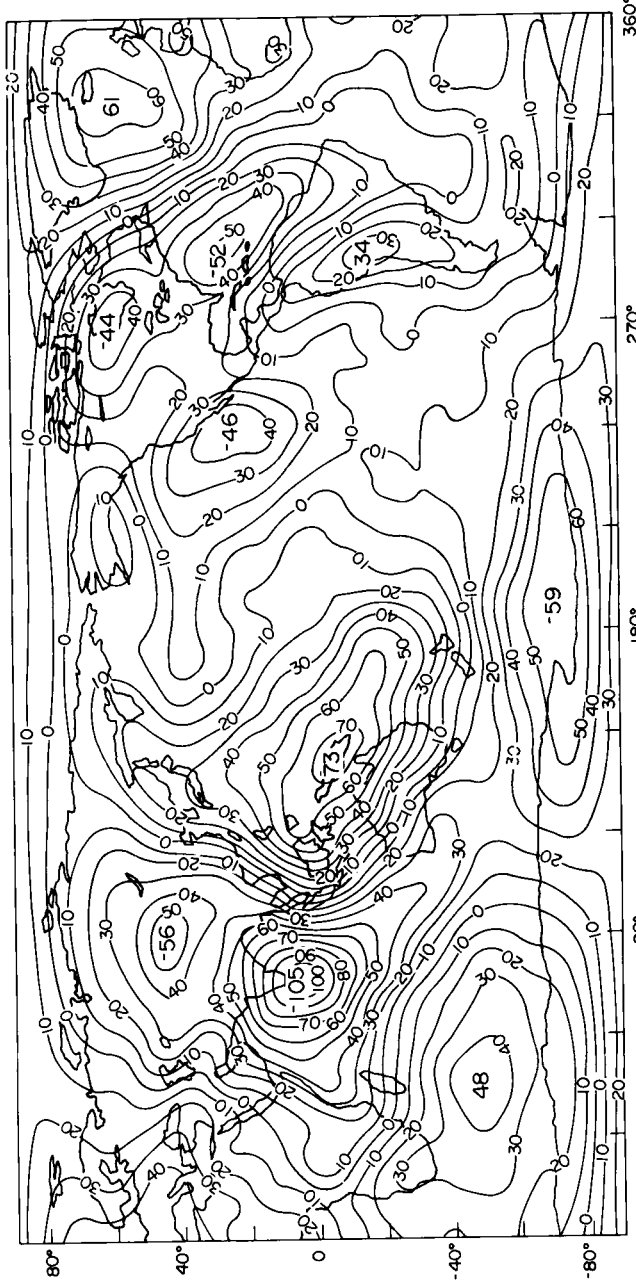


FIG. 1.11 Geoid surface computed from GEM 10 model. Heights in metres above a mean spheroid of flattening 1/298.255. Source: King-Hele (1978). This world map is based upon the Plate Carrée or Cylindrical equidistant projection (No. 2) in the normal aspect. The full graticule is not shown, but marginal ticks indicate 20° values for latitude and 30° values for longitude.

than is a plane. On the other hand, the list is in order of increasing mathematical difficulty. The formulae needed to define position, to determine the relationships between distances and angles on a plane are simpler than are those for the curved surface of a sphere. These, in turn, are simpler than the corresponding formulae for an ellipsoid. Bearing in mind the desirability of using the simplest reference figure which is compatible with accuracy of representation, it follows that we should inspect the properties of each kind of reference surface to discover when it should be used.

The plane reference surface

At first sight it may seem to be a retrograde step to assume that the earth is a plane. However, it is a very useful assumption because it is so simple to use. For a start we can avoid the whole problem of map projection transformations which are the preoccupation of this book. Figure 1.02 indicates that near a point A on the curved surface of the earth, the tangent to the curved surface also lies close to it. The tangent plane and the curved surface only diverge from one another as one moves away from A. It may therefore be argued that if we only need to make a survey of a small area around A, it is reasonable to assume that we are making the measurements on the tangent plane. The survey can be computed by the methods of plane trigonometry (it is then called *plane surveying*). Plotting of the map can be done simply by converting ground dimensions to the required map scale. The crux of the argument is the definition of what is represented by the immediate vicinity around the point A. It implies that the plane assumption should be confined to the preparation of maps of small areas, but it still remains necessary to define what we mean by a small area. We defer quantitative consideration of this problem until Chapter 15 (pp. 310–335) because it is desirable to consider this assumption together with the kinds of map projections which are used by surveyors, and which are also important in large-scale cartography of small areas.

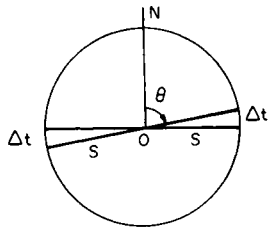
The spherical assumption

We have already commented upon the fact that, at a scale of 1/100 000 000, the lengths of the two axes of the spheroid differ by about the width of the lines needed to draw them. This implies that the main use of the spherical assumption will occur in the preparation and use of comparatively small format maps showing large parts of the earth's surface such as maps of the world, a hemisphere, a continent or even a large country, such as appear in atlases. The question to be answered is:

'What is the approximate maximum scale at which the spherical assumption can be justified?'

This subject was tackled theoretically by Driencourt in 1932, and his work has been reproduced more recently by Richardus and Adler (1972). Therefore we need not reproduce the detailed mathematical argument here. Driencourt showed that the largest errors occur in lines which are orientated east or west from a point, and that the maximum linear displacement, Δt is directed northwards or southwards. He calculated the following results for a line of length s (km). The following table shows that the discrepancy Δt at a distance of approximately 100 km from the central point, does not exceed 1 mm or 10^{-8}. At a distance of 1000 km from the point the proportion $\Delta t/s$ is approximately 10^{-5}, which is about three times the present precision of electronic distance measurement.

Tobler (1964) also investigated the problem from the point of view of mapping the United States of America. He calculated the distances and bearings between 200 randomly selected places in the USA for both the sphere and spheroid. He used the Clarke 1866 Figure of the Earth, which



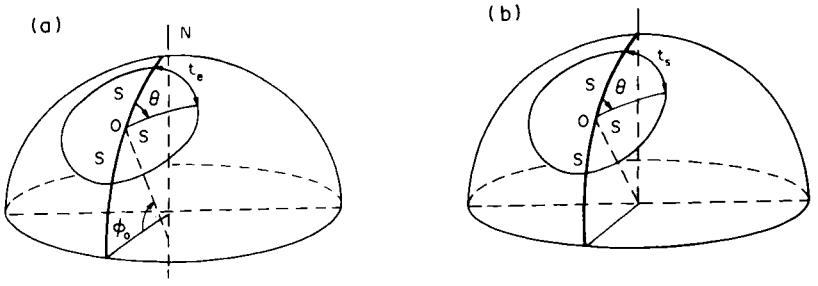


FIG. 1.12 The geometry of the reference surfaces; a comparison between the spheroidal (a) and spherical (b) surfaces showing corresponding observations. A line of length s is measured from the point *O* along a bearing *θ*. On the spheroid this bearing traces the arc t_e at the distance s from *O*; on the sphere the corresponding arc is t_s. The difference $\Delta t = t_e - t_s$ in these arc lengths is a measure of the discrepancies which occur if the earth is assumed to be spherical. The amount varies with the size of the angle *θ* and with the distance s. (Source: Driencourt and Laborde, 1932.)

TABLE 1.02 *Driencourt's values for the maximum discrepancies between the lengths of great circle arcs and geodesics*

s (km)	103·7	184·4	327·9	583·1	1039·9
Δt (m)	0·001	0·01	0·1	1·0	10·0

was that still in use for North America at the time. For the spherical assumption he chose as the radius R = 6378·206 km, which is the equatorial radius for the Clarke 1866 figure. The results are given in Table 1.03.

If we assume 0·2 mm to be the smallest linear distance which can be measured on a map without special magnification, and if we take Tobler's average difference in distance as being equal to this, then the largest scale at which the USA might be represented by a projection of the sphere is 1/370 000. However, the spread of the results, characterised by the values for the standard deviation and the two extremes, indicates that it would be optimistic to use the spherical assumption at such a large scale and imagine that no errors in mapping would arise from this cause. The figures suggest that, strictly speaking, the spherical assumption ought to be confined to use for maps of scale 1/15 000 000 or smaller, which is about the scale at which 7·8 km is represented by 0·2 mm. In practical cartography, however, the limit of using the spherical assumption is usually taken to be a scale of 1/5 000 000 or thereabouts. Using Tobler's data it can be argued that at this scale about two-thirds of the points lie within 1 mm of the spheroidal position if mapped on a sphere. We shall see later that this discrepancy is small compared with the displacements which are inherent in the process of representing a large country at a small scale on a plane map.

A third approach has been adopted by Snyder (1987a), who has applied the same distortion theory which we shall investigate in the study of plane map projections to the projection of the spheroid to the sphere. This gives rise to a series of values for *particular scales* and distortion characteristics

TABLE 1.03 *Comparison of the differences in distances and bearing between 200 points in the United States of America computed on both a sphere and a spheroid*

	Distance difference (km)	Angular difference (degrees)
Average	0·074	0·006
Standard deviation	3·053	0·083
Minimum	−6·100	−0·150
Maximum	+7·844	+0·159

which are introduced in Chapter 5. The numerical characteristics thus obtained may be used to determine the maximum scale at which the distortion cannot be recognised on a map.

The spheroidal assumption

Obviously the spheroid fits the shape of the geoid more closely than does a sphere. Consequently this is the reference surface which ought to be employed in surveying. This is because the survey of a country is first computed to determine the positions of the control points in their natural dimensions or, as it were, for a map of scale 1/1. Consequently the small discrepancies in position (or *closing errors*) may be expressed to the nearest millimetre or less on the ground and not absorbed by scale reduction as would happen if the results of a survey were first plotted on a sheet of paper. In order to appreciate the quality and precision of the work it is desirable to make these computations with respect to a particular reference spheroid rather than risking the introduction of errors arising from assuming a flat or spherical earth. At the later stage of producing topographical and other map series, extending throughout an entire country, continuity of information across boundaries of adjacent map sheets is important. Hence it is desirable to use the reference ellipsoid as the basis of such maps. It is also used for the compilation of large-scale navigation charts and small-scale charts to the approximate limit of 1/4 000 000–1/5 000 000.

Table 1.01, on page 10, indicated that about 15 different reference ellipsoids may be encountered in world mapping, and about six of them are in common use. From the point of view of practical cartographic work the correct spheroid for use should always be clearly stated in the mapping specification. From the point of view of evaluating existing topographical or other maps as source documents for compilation, references such as the United Nations' summaries on the status of world topographic mapping (United Nations, 1970, 1976, 1979) and the national survey reports provide the information which is needed. In an analysis of the UN data Brandenberger and Gosh (1985) have estimated that nearly 93% of the earth's land area has been mapped on only four of the classical figures. These are:

International spheroid	28·3%
Krasovsky spheroid	25%
Bessel spheroid	19·9%
Clarke 1880 spheroid	19·4%.

Originally a particular spheroid was selected by the national survey because the parameters of the figure fitted the observed data better than any other. A typical example of this was the use of the Airy spheroid for

Great Britain, for this had been derived from astro-geodetic distances obtained during the original Primary Triangulation of the country. In the days before digital computing, once a national survey had been computed using a particular reference figure it would have been extremely inconvenient and costly to convert the positions of many hundreds or even thousands of control points to another spheroid. It was done in the USSR when the decision was taken in 1942 to transform the entire control network from the Bessel spheroid to the newly described Krasovsky figure, but that was a practically unique example. It follows that usually a national survey continued to be based upon a particular figure long after the original reasons for its choice had ceased to be valid.

This argument carries less weight today than before digital computing became commonplace. It is interesting to note in this context that probably the first major use of digital computing in geodesy and surveying was the work undertaken by the US Army Map Service shortly after World War II, when they accomplished the formidable task of reducing the national surveys of western Europe to a common datum on the International Spheroid. This is known as the *European Datum, 1950*, or ED50. This network had hitherto been based upon a multiplicity of different points of origin, reference spheroids, units of measure and projections. We shall also refer to the change in the North American Datum from NAD 27 into NAD 83 during the 1980s, which amongst other changes includes that from the Clarke 1866 figure to GRS 80.

Nevertheless the use of different figures still remains. It arises partly from historical accident, partly from inertia and partly for reasons of national prestige. Sometimes it also happens that the chosen spheroid fits the shape of the geoid in that country better than any of the others.

Finally the continuity of use is important. Indeed Chovitz (1981) has argued that this continuity is at least as important as the formal accuracy of recording the length of the major semi-axis and flattening. Some of the better-known figures, such as Airy, Everest and the three useful Clarke determinations, have been slightly modified on many occasions for use in different places or for different purposes. Typical examples include retaining the original value for the semi-axis, a, but using it with a slightly different (rounded) value for f. Other changes have been enforced by the discrepancies introduced to the dimensions of the semi-axes through converting from British Standard into metric units or vice-versa. For example, Strasser (1975) has shown how US legislation concerning the definition of the metre has created numerous difficulties in reconciling different versions of the Clarke 1866 figure. Sometimes we know enough about the history of a survey to understand where discrepancies have arisen. More often it may be extremely difficult to reconcile these so that mistakes are sometimes made in choosing the correct version of Everest or Airy.

CHAPTER 2

Coordinate reference systems on the plane

> It is impossible not to feel stirred at the thought of the emotions of men at certain historic moments of adventure and discovery – Columbus when he first saw the Western shore, Pizarro when he stared at the Pacific Ocean, Franklin when the electric spark came from the string of his kite, Galileo when he first turned his telescope to the heavens. Such moments are also granted to students in the abstract region of thought, and high among them must be placed the morning when Descartes lay in bed and invented the method of co-ordinate geometry.
>
> A. N. Whitehead

Introduction

In this chapter we review some of the fundamental ideas about the plane coordinate systems which are used in surveying and mapping, both from the viewpoint of studying the mathematics of map projections and the practical tasks which arise in cartography.

Coordinates are a convenient method of recording position in space. They may be used to locate position in two dimensions, such as a point on a graph. An extension of this method to map use allows the location of a place by its *grid reference*. Definition of coordinate position on the surface of a three-dimensional body such as a sphere or spheroid is rather more difficult. However, the reader should already be aware of the method of describing location by means of latitude and longitude, which are *geographical coordinates*. These are defined in Chapters 3 and 4, where the differences between defining latitude on a sphere and on a spheroid are introduced. In addition to providing a means of reference, coordinates can also be used as a convenient way of solving certain geometrical problems. The branch of mathematics known as *coordinate geometry* analyses problems through the relationship between points as defined by their coordinates. By these means, for example, it is possible to derive algebraic expressions defining different kinds of curve which cannot be done by Euclidean geometry. Coordinate geometry is an exceptionally powerful tool in the study of the theory of map projections, and without its help it is practically impossible to pass beyond the elementary descriptive stage. Plane coordinate geometry is usually studied first through the

medium of the *conic* sections or the definition of the different kinds of curve formed by the surface of a cone where this has been intersected by a plane. Two of the resulting sections, the ellipse and the circle, are of fundamental importance to the theory of distortions in map projections.

There are an infinite number of ways in which one point on a plane surface may be referred to another point on the same plane. Every map projection creates a unique reference system which satisfies this requirement and an infinity of different map projections could theoretically be described. However it is desirable to use some kind of coordinate system to describe, analyse and construct each of these projections. Any system to be used for such purposes ought to be easy to understand and simple to express algebraically. For plane representation the choice lies between *plane cartesian coordinates* and *polar coordinates.*

Plane cartesian coordinates

The reader will already be familiar with graphs as a method of plotting two variables on specially ruled paper and with the *National Grid* on Ordnance Survey maps. The graph and the National Grid are simple, but special, examples of plane cartesian coordinates. In the general case, *any* plane coordinate system which makes use of linear measurements in two directions from a pair of fixed axes can be regarded as a cartesian system. The coordinate system comprises sets or *families* of lines which intersect one another to form a *network* when plotted. The only necessary conditions which must be fulfilled are:

- that the two families of lines are distinct from one another;
- that every line of one family should intersect every line of the other family at one point only;
- that no two lines of the same family should intersect one another.

Thus a cartesian coordinate system can comprise families of straight lines or curves which may intersect at any angle. However, it is a distinct advantage if the special case is chosen in which both families of lines are straight and that they are *orthogonal*, or intersect at right angles. This special case, characterised by ordinary graph paper and by the National Grid on Ordnance Survey maps, may be called a *plane rectangular cartesian coordinate system*, or, in short, *rectangular coordinates.*

In Fig. 2.01 the *origin* of the rectangular coordinate system is the point O, through which two orthogonal axes, OX and OY, have been plotted. These axes define the directions of the two families of lines. Since the axes are straight lines and perpendicular to one another, it follows that all the lines composing one family will be parallel to one another and that all the points of intersection within the network are made from lines which are perpendicular to one another. The position of a point A is defined by

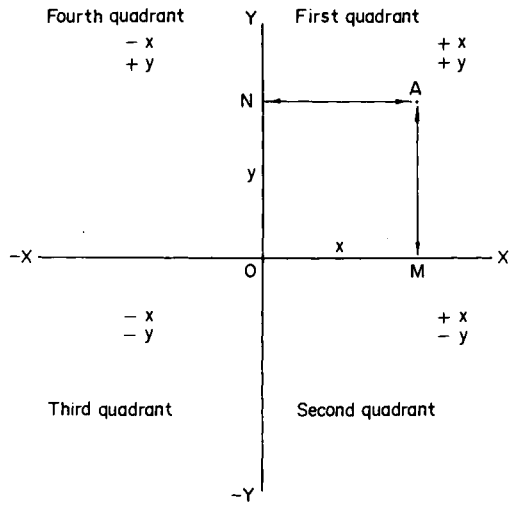


FIG. 2.01 Plane rectangular cartesian coordinates.

the two linear measurements OM and ON made from the origin to the points M and N on the two axes, which are drawn perpendicular from A to the axes. Clearly AM is parallel to OY and AN is parallel to OX. The mathematical convention is to refer to the horizontal axis OX as the X-axis or *abscissa*. The vertical line OY is called the Y-axis or *ordinate*. However, the convention is not always observed in the study of geodesy, surveying and map projections. In some books the notation is reversed and OX is the axis pointing upwards on the page. There are cogent reasons for this change in notation, to do with the direction in which angles are measured, as described on p. 34, but the change in axes is extremely confusing to the beginner. We shall use the standard mathematical, or graph, convention throughout most of this book and refer to the coordinates of the point A as being (x, y) according to the axes illustrated in Fig. 2.01. It is not until Chapter 15 that we have to change the notation for particular purposes. Even then we use it sparingly.

The units into which the axes are subdivided for the purposes of linear measurement are quite arbitrary. For example, graph paper is available with both millimetre and inch ruling, with various combinations of multiples and fractions of these. The National Grid is measured in metres. We shall make considerable use of units of earth radius, R, in which coordinates are expressed in multiples or decimals of R without having to convert into units suitable for plotting on a sheet of paper.

There is a sign convention to be observed in the use of rectangular coordinates. This states that the X-axis is reckoned positive towards the right and the Y-axis is positive towards the top of the page. In other words, a point in the top right-hand quarter of a graph illustrated by Fig.

2.01 is defined by positive values of x and y, whereas a point in the bottom left-hand quarter has negative values for x and y. The quarters are termed *quadrants* and these are numbered 1–4 in a clockwise direction commencing with the top right quadrant. Hence the sign convention is:

1st quadrant	+x,	+y
2nd quadrant	+x,	−y
3rd quadrant	−x,	−y
4th quadrant	−x,	+y

The map grid as an example of plane rectangular coordinates

A grid has been defined in the *Glossary of Technical Terms in Cartography* (Royal Society, 1966) as * *'a cartesian reference system using distances measured on a chosen projection'*.† In the first edition of this book the author disagreed with the last seven words in this definition, but as a major contributor to the Glossary felt a certain loyalty to the deliberations of the working group, limiting himself to making only a mild criticism of this particular definition. Professor E. H. Thompson (1973) was not restrained by such inhibitions, and in his important review of the first edition of this book made the following characteristically forthright statement:

> It is sad to see an author, who has clearly thought out so much of the problem for himself, committing old faults because his courage fails him at the last minute. He says 'For the moment it will suffice to regard a grid as a system of rectangular coordinates superimposed upon a plane corresponding to the ground'. Why 'For the moment'? Grids are simply sets of squares and to paraphrase Gertrude Stein, a square is a square is a square. It is indeed a pity that we are also given a definition from the Royal Society *Glossary of Technical Terms in Cartography*.... Whatever has a projection to do with a grid? The sin is Dr Maling's only in so far as he perpetuates it and he barely does that for he says, about the above definition, '... the last seven words... are probably necessary but tend to confuse the issue'. They are not necessary and they do indeed confuse the issue by being quite wrong.

One family of lines is orientated approximately north–south and the other family, by definition, is perpendicular to them. Measurements along the axes are made in some units used for ground measurement. Nowadays the metric system is used almost everywhere, but formerly some grids used feet or yards as the unit. By virtue of the approximate orientation

†Frequent reference will be made in this book to the labours of the United Kingdom Working Group on Terminology and to the preparation of the *Glossary of Technical Terms in Cartography*, published by the Royal Society in 1966. The definitions in that work were subsequently combined with other national contributions to the *Multilingual Dictionary of Technical Terms in Cartography*, published by ICA in 1973. The preferred terms relating to map projections which appear in those works are used throughout the book. Definitions which are those used in the *Glossary* are prefaced with the symbol *.

of a grid, the abscissa of a point is usually called its *Easting* and the ordinate is its *Northing*. Thus E corresponds to x and N corresponds to y in the mathematical and graph conventions. We will introduce this substitution without further comment where it is appropriate to refer to a point by its (E, N) coordinates rather than by (x, y). The order in which the grid coordinates are recorded is often confusing to the beginner, who has probably only just learnt to describe geographical position in the order 'latitude-followed-by-longitude'. If it is remembered that a grid is like a graph, then the logic of using 'Easting-followed-by-Northing' matching the 'x-followed-by-y' graph convention is apparent.

We do not attempt to describe in detail how a grid reference may be obtained from a map, for it is assumed that the reader can do this already. Military manuals, such as Ministry of Defence (1973, 1978) are always painstaking in describing this aspect of map use, for it is vital to military communications. The practices adopted by the Ordnance Survey for use with the National Grid are described in Ordnance Survey (1951) and Harley (1975). This distinguishes the slightly different procedures to be adopted at different map scales. Moreover many Ordnance Survey and other national survey maps have the appropriate instructions, with a worked example, printed in the margin.

Because a grid is a form of graph it must have an origin. Moreover if the grid is to satisfy its purpose to serve as a national or international standard of reference, the point of origin must be explicitly stated, together with the orientation of the axes at this point. It is this aspect of a grid which introduces the confusing ideas in the second part of the definition given on p. 30. For example, the National Grid (Fig. 2.02) has its origin at the point with latitude 49°N, longitude 2°W. This is situated in the Golfe de St Malo, about 20 km south-east of St Helier in Jersey. The same point is also taken as the origin of the map projection used by the Ordnance Survey for all topographical maps of England, Scotland and Wales. We defer the projection part of the problem to a later chapter. Here it is desirable to consider two properties of the grid, its orientation and the system of numbering along the axes.

The ordinate of the system is orientated so that it coincides with the meridian 2°W. It follows that since all meridians point towards true north (see Chapter 3 for justification of this statement), the ordinate of the National Grid also points towards true north. Since the grid is composed of families of straight lines, it follows that all other vertical grid lines point in the constant direction defined by the ordinate. This constant direction may be called *grid north*. On the other hand all meridians converge towards the geographical poles, therefore a meridian through a point lying east or west of longitude 2°W does not coincide everywhere with a grid line through the same point. This gives rise to the angular discrepancy between the meridians and grid lines which is illustrated

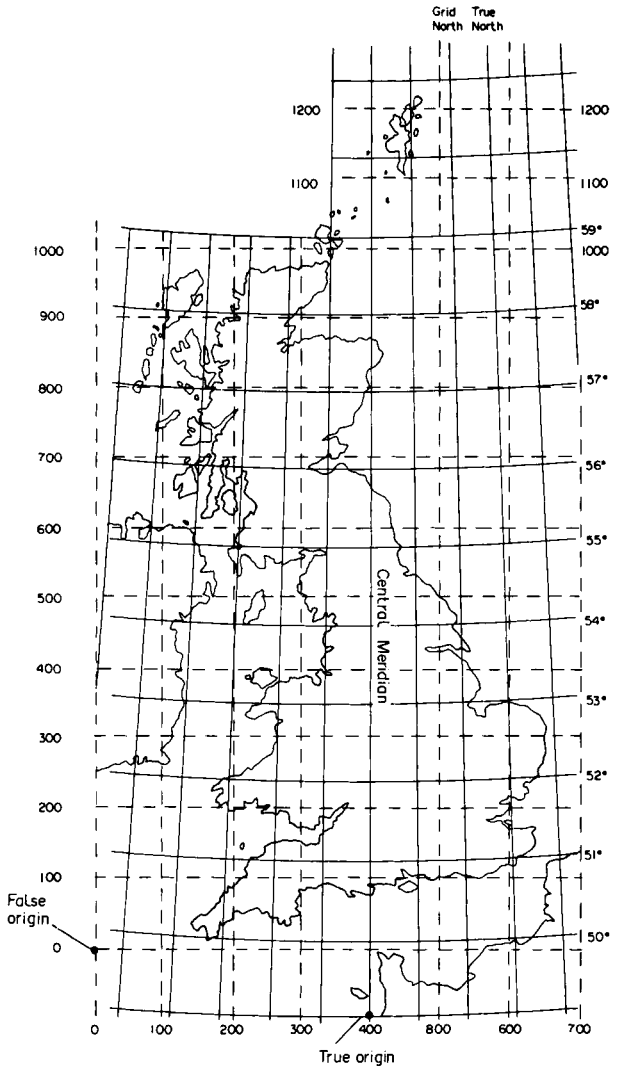


FIG. 2.02 The National Grid of Great Britain, showing the relationship between the grid lines (broken) at every 100 km, and the graticule of meridians and parallels (full lines) at 1° intervals of latitude and longitude.

in a much exaggerated form in Fig. 2.03. The angle is known as *grid convergence*. Within the range in longitude occupied by southern England, the amount of convergence is small, for example it is 2° 54′ near Lands' End and nearly 3° on the Norfolk coast.

The choice of the meridian 2°W as the longitude for the origin is simply because this lies near the middle of the part of the British Isles covered by the National Grid. It is a line which passes through the Isle of Purbeck

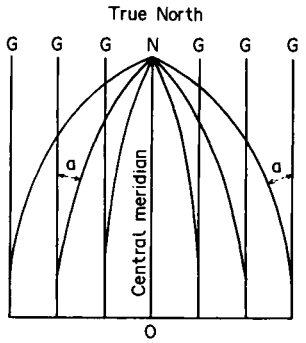


FIG. 2.03 The relationship between true north, represented by the meridians converging to the geographical pole, N, and grid north which is a constant direction for any particular grid. Grid convergence is indicated by the angles, α.

in Dorset, through Birmingham, Berwick and Fraserburgh. From the sign convention used with graphs this means that everywhere in Britain lying to the west of the Birmingham–Berwick–Fraserburgh line, i.e. all Wales, most of Scotland and much of England, would be assigned negative Easting coordinates and referred to in this inconvenient way. The method of overcoming likely confusion is to imagine that the origin of the National Grid has been shifted westwards until the whole country lies in the first quadrant of the graph. In the example of the British National Grid the shift in origin is 400 km to the west and 100 km to the north of the point near the Channel Islands, so that zero on the National Grid lies at a point located about 80 km west of the Scilly Isles. This is equivalent to assigning the arbitrary coordinate values $E = 400\,000$ m, $N = 100\,000$ m to the true origin and renumbering the grid lines. The point $E = 0$ m, $N = 0$ m is referred to as the *false origin* of the grid to distinguish it from the point in latitude 49°N, 2°W which is the true origin. They way in which the shift has been applied may be imagined mathematically as the parallel shift of each axis through the defined distances. This is called *translation of the axes.*

Plane polar coordinates

Polar coordinates define position by means of one linear measurement and one angular measurement. The pair of orthogonal axes passing through the origin is replaced by a single line OQ, in Fig. 2.04, passing through the origin O, or *pole* of the system. The position of any point A may be defined with reference to this pole and the *polar axis* or *initial line*, OQ by means of the distance $OA = r$ and the angle $QOA = \theta$. The line OA is known as the *radius vector* and the angle θ is the *vectorial angle*

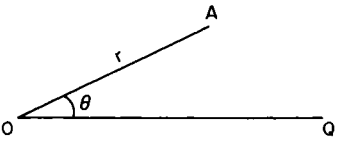


FIG. 2.04 Plane polar coordinates.

which the radius vector makes with the initial line. Hence the position of A may be defined by the coordinates (r, θ). The order of referring to the radius vector followed by the vectorial angle is standard to all branches of pure and applied mathematics. The vectorial angle may be expressed in sexagesimal (degree) or centisimal (grad) units to plot or locate a point instrumentally.*

In the theoretical derivation of map projections, where θ enters directly into an equation and is not introduced as some trigonometric function of the angle, it is necessary to express this angle in *absolute angular units*, or radians. This is because both elements of the coordinate system must have the character of length.

The direction in which the vectorial angle is measured depends upon the purpose for which polar coordinates are used. Usually the mathematician regards $+\theta$ as the *anticlockwise angle measured from the initial line.* This is the sign convention which is used, for example, in vector algebra. On the other hand, the navigator, surveyor and cartographer are accustomed *to measure a positive angle in the clockwise direction.* This is because direction on the earth's surface is conventionally measured clockwise from north or clockwise from a reference object. In many practical applications, formal recognition of the sign of an angle is unimportant because the user can visualise the relationship between angles measured on the 360° circle. However, difficulties arise in automatic data processing because the standard subroutines, for example those to convert from rectangular into polar coordinates, invariably use the mathematical convention. This kind of calculation, which is described in the next section, is extremely common in surveying and cartography. Consequently the user of a computer or calculator must be aware of the difference in convention, how the instrument deals with such data and write suitable program steps which overcome the difficulty. Similarly in writing programs for digital processing it is frequently necessary to introduce a series of tests and conditional statements to allow uninterrupted processing of data which have been collected according to the clockwise convention. The simplest way of overcoming the difficulty is to interchange the axes, so that the x-axis points towards the north. This is equivalent to a rotation

*One right angle is representated by 90° in sexagesimal notation, 100^g in centisimal units or $\pi/2$ radians. Many pocket calculators can operate in all three modes.

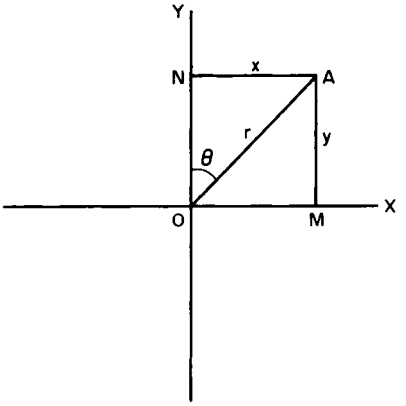


FIG. 2.05 The relationship between plane rectangular and plane polar coordinates with common origin and one common axis.

plus a reflection of Fig. 2.05, which may be verified by tracing this diagram on a piece of transparent plastic.

Transformation from polar into rectangular coordinates and vice-versa

Figure 2.05 illustrates the relationship between the rectangular and polar coordinates of a point A. The rectangular coordinates of the point are (x, y) referred to the origin O and the axes OX and OY. Superimposed upon this is a system of polar coordinates in which the pole also lies at O and the initial line coincides with OY. Then the polar coordinates of A are (r, θ) where $r = OA$ and $\theta =$ angle YOA, $AN = x$ and $AM = NO = y$. It is evident from the right-angled triangle AON that

$$x = r \,.\, \sin\theta \tag{2.01}$$

$$y = r \,.\, \cos\theta \tag{2.02}$$

The inverse transformation from rectangular to polar coordinates can be accomplished using a variety of different formulae. For example

$$\tan\theta = x/y \tag{2.03}$$

$$r = y \,.\, \sec\theta \tag{2.04}$$

$$r = x \,.\, \operatorname{cosec}\theta \tag{2.05}$$

$$r^2 = x^2 + y^2 \tag{2.06}$$

$$\sin\theta = x/r \tag{2.07}$$

$$\cos\theta = y/r \tag{2.08}$$

Note that these expressions are based upon the assumption that the angle θ has been measured 'clockwise-from-grid-north'. The coordinate expressions corresponding to these in most mathematical textbooks are derived from the complement of the vectorial angle, i.e. $AOX = 90° - \theta$.

From the expressions which may be used to transform from rectangular to polar coordinates, the formulae (2.03) and either (2.04) or (2.05) used to be the most convenient in numerical work, and the reader would be warned against using Pythagoras' Theorem (2.06) to find the length of the radius vector because this was slow and inconvenient to calculate by logarithms. Nowadays most pocket calculators can be used to obtain square roots directly, so this caveat no longer applies.

Two-dimensional coordinate transformations

A series of numerical procedures which are commonly required in the mapping sciences are the two-dimensional linear transformations from one cartesian coordinate system into another. We provide here five examples of applications, and this list is by no means exhaustive. It includes:

(1) Determination of the positions of intersections of a grid to be plotted on a map manuscript which has been compiled from and shows a different grid. This is necessary for mapping the zone of overlap between two grid systems and both of them have to be shown on the map.

(2) Determination of the positions of intersections of a new grid to be plotted on a map manuscript originally compiled on a different grid which has been superseded. Now that most national surveys are based upon either the Universal Transverse Mercator (UTM) projection or the similar Soviet Unified Reference System (SURS), the need for this conversion is much less than it was in the early postwar decades, when many separate, or local, grid systems were still in use.

(3) Conversion of the coordinate output of some other mapping process so that the results can be used with a particular grid. A typical example of this kind of work is when aerial triangulation has been carried out in an analogue photogrammetric plotter. The output from this includes a stream of (X, Y) *model coordinates* for control points which have been observed in the plotter and whose positions are recorded with respect to the axial movements of the plotter. These now have to be transferred to the same system as the map grid in order to fit the photogrammetric control to ground surveys. The concept of the *analytical plotter* which has more or less replaced the older analogue instruments is based upon continuous transformation from the plane of the aerial photograph to that of the map by digital methods.

(4) Perhaps the most important application of all now arises in digital mapping, in the use of *vector* digitised map information to refer digitiser coordinates to the map grid. The majority of instruments functioning in

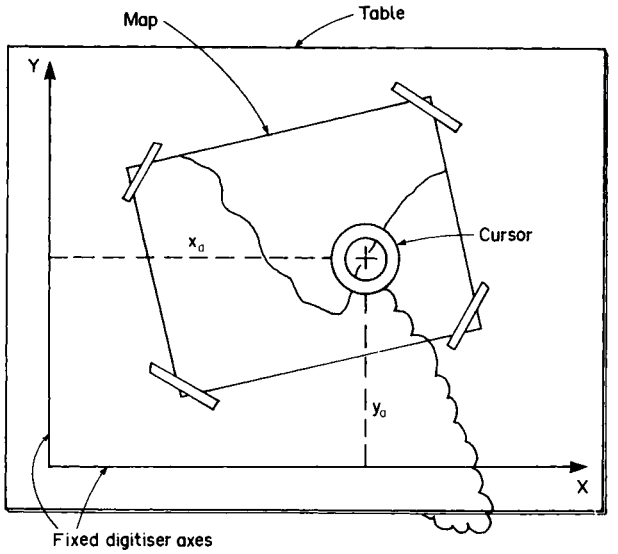


FIG. 2.06 Diagrammatic representation of a vector digitiser used to measure and output the coordinates of the position of the cursor. (Source: Maling, 1989.)

the vector mode comprise a special table containing the electronic hardware which converts the positions of a measuring mark mounted in a special cursor into rectangular coordinates defined by the manufacturer of the table. Information about position is obtained by pointing to or tracing the map detail (called *line-following*) with the measuring mark. The coordinates of a single point or points along a line are recorded and stored in digital form on tape or disc according to the (x, y) coordinate system built into the instrument. Hence the (E, N) grid of a topographical map is converted into the (x, y) coordinates of the digitiser and the precise relationship between the two depends upon the way in which the map sheet was placed upon and attached to the table. In order to reproduce any of the map detail in a desired form it is necessary to convert back from the (x, y) system of digitised coordinates into the (E, N) system of the map grid. This is usually done by digitising the four corners of the map and using these control points to determine the translation, rotation and scale change components of the transformation.

(5) A second stage of this kind of digital mapping is contained in the need to change from one map projection to another, from a source map on one map projection to a new map which is compiled upon another. We consider this particular application of the two-dimensional transformations in detail in Chapter 19. Here we confine our attention to the two simplest methods:

- *The linear conformal, similarity or Helmert transformation*, expressed in the general form:

$$\left.\begin{aligned} X &= A + Cx + Dy \\ Y &= B - Dx + Cy \end{aligned}\right\} \qquad (2.09)$$

- The affine transformation:

$$\left.\begin{aligned} X &= A + Cx + Dy \\ Y &= B - Ex + Fy \end{aligned}\right\} \qquad (2.10)$$

In these equations the known (x, y) coordinates of a point in one system are transformed into the (X, Y) coordinates of a second system, through the use of four or six coefficients A–F. In the first we see that the C and D coefficients are common to both the equations for X and Y, but in affine transformation it is necessary to introduce separate corrections for each direction. The risk of confusion of the coefficient E in equation (2.10) with the abbreviation for Easting should be noted.

Linear conformal, similarity or Helmert transformation

Both transformations may be resolved into three components:

- translation of the axes or change of origin, corresponding to the coefficients A and B in both equations (2.09 and 2.10);
- change in scale from one grid system to the other;
- rotation of the axes of one grid system with respect to their directions in the other.

The difference between the Helmert and affine transformations comes in the treatment of scale changes and rotations of the axes.

Translation of the axes or change of origin

We have already described this transformation for it has been used to introduce a false origin to a grid. This is simplest if the axes of the original system and those of the final system are parallel to one another as illustrated in Fig. 2.07. In this figure the point A has (x, y) coordinates in the original system which has its origin at O. We wish to refer the point to the second system in (x′, y′) coordinates which have their origin at O'. The differences between O and O' are the coordinate displacements x″ and y″. It follows that the new coordinates of A may be written

$$x' = x + x'' \qquad (2.11)$$

$$y' = y + y'' \qquad (2.12)$$

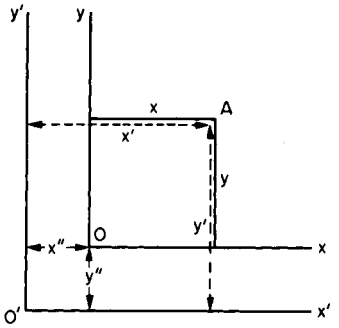


FIG. 2.07 Translation of the axes of a plane rectangular coordinate system.

The signs of x″ and y″ depend upon the direction in which the shift has been made. However, in dealing with grids of topographical maps, the false origin has usually been assigned to a position which lies to the south and west of any point likely to be referred to the grid, thereby avoiding the inconvenience of having negative grid references. It follows that normally $x' > x$ and that $y' > y$ so that x″ and y″ are both positive corrections. We may express the pair of equations (2.11) and (2.12) in the form of matrix addition,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x'' \\ y'' \end{pmatrix} \tag{2.13}$$

Change in scale from one coordinate system to another

Consider two points, *A* and *B*, which are common to two coordinate systems. In the first system the straight line *AB* joins the pair of points and in the second system the corresponding line is *ab*. If $AB \neq ab$, a scale factor must be introduced to convert coordinates in the first system into coordinates within the second system. This scale factor is

$$m = ab/AB \tag{2.14}$$

from which it follows that

$$x' = m \,.\, x \tag{2.15}$$

$$y' = m \,.\, y \tag{2.16}$$

In matrix notation this has the form

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = m \,.\, \begin{pmatrix} x \\ y \end{pmatrix} \tag{2.17}$$

where the term m is appropriately called a *scalar*. A typical application

of this part of the transformation is the conversion of (x, y) *projection coordinates*, which are given in units of earth radius, into the (x′, y′) system of *master grid coordinates* which are needed to plot points on a master grid in millimetres. We shall see in Chapter 8 that this is the customary method of constructing a map to a required scale.

Rotation of the axes about the origin

We assume that the origin of each system is the same point, O, but the axes have been rotated through the angle α. Thus OX becomes OX' and OY becomes OY', as illustrated in Figs 2.08 and 2.09. These two figures illustrate the difference between the clockwise and anticlockwise rotations of the axes. We shall study the effect of a clockwise rotation of the axes in detail.

If $A = (x, y)$ in the first system it is required to determine its (x′, y′) coordinates after rotation of the axes to form the second system. From equations (2.01) and (2.02) we know that $x = r.\sin\theta$ and $y = r.\cos\theta$, where θ is the angle AOY. Moreover the angle $AOY' = \theta - \alpha$. Therefore

$$x' = r.\sin(\theta - \alpha) \tag{2.18}$$

$$y' = r.\cos(\theta - \alpha) \tag{2.19}$$

The sine and cosine of the difference between two angles are well-known formulae from plane trigonometry. Here

$$\sin(\theta - \alpha) = \sin\theta.\cos\alpha - \cos\theta.\sin\alpha \tag{2.20}$$

$$\cos(\theta - \alpha) = \cos\theta.\cos\alpha + \sin\theta.\sin\alpha \tag{2.21}$$

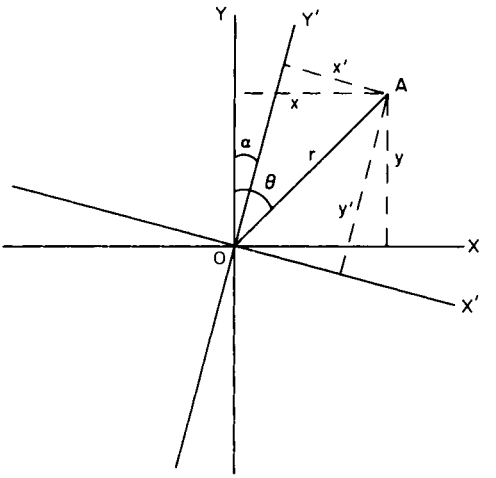


FIG. 2.08 Clockwise rotation of plane rectangular coordinate axes about the origin.

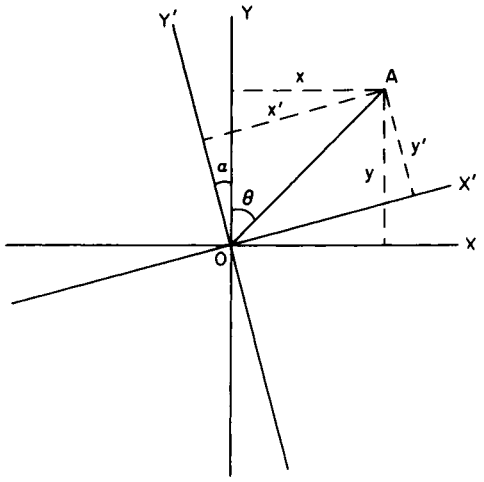


FIG. 2.09 Anticlockwise rotation of plane rectangular coordinate axes about the origin.

Substituting these expressions in equations (2.18) and (2.19)

$$x' = r \,.\, \sin\theta \,.\, \cos\alpha - r \,.\, \cos\theta \,.\, \sin\alpha \tag{2.22}$$

$$y' = r \,.\, \cos\theta \,.\, \cos\alpha + r \,.\, \sin\theta \,.\, \sin\alpha \tag{2.23}$$

From equations (2.01) and (2.02) we may now substitute x and y for $r \,.\, \sin\theta$ and $r \,.\, \cos\theta$ respectively. Thus

$$x' = x \,.\, \cos\alpha - y \,.\, \sin\alpha \tag{2.24}$$

$$y' = x \,.\, \sin\alpha + y \,.\, \cos\alpha \tag{2.25}$$

Note the order in which the terms for x and y are written. This corresponds to the rules governing the order in which terms and coefficients are written in matrices, so that these two equations have the matrix notation

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} . \begin{pmatrix} x \\ y \end{pmatrix} \tag{2.26}$$

The 2×2 matrix containing the trigonometric coefficients is known as the *rotation matrix*. We turn now to the anticlockwise rotation of the axes illustrated by Fig. 2.09, where the angle $Y'OA = \theta + \alpha$. Using the same arguments with the trigonometric expressions defining the sine and cosine of the sum of two angles, the final equations are:

$$x' = x \,.\, \cos\alpha + y \,.\, \sin\alpha \tag{2.27}$$

$$y' = -x \,.\, \sin\alpha + y \,.\, \cos\alpha \tag{2.28}$$

which means that the rotation matrix is now

$$\mathbf{R} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \tag{2.29}$$

We observe that the two elements in $\sin x$ have different signs and the position of $-\sin\alpha$ has changed between (2.26) which refers to the clockwise rotation and (2.29) describing the anticlockwise rotation.

Coordinate transformations involving all three displacements

We may now combine the effects of all three displacements to produce the pair of equations

$$x' = (m \,.\, x \,.\, \cos\alpha + m \,.\, y \,.\, \sin\alpha) + x'' \tag{2.30}$$

$$y' = (-m \,.\, x \,.\, \sin\alpha + m \,.\, y \,.\, \cos\alpha) + y'' \tag{2.31}$$

Several different versions may be used to express the result in matrix form. The simplest is to write

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} m \,.\, \cos\alpha & m \,.\, \sin\alpha \\ -m \,.\, \sin\alpha & m \,.\, \cos\alpha \end{pmatrix} . \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x'' \\ y'' \end{pmatrix} \tag{2.32}$$

In many survey applications there is a convention of writing $P = m \,.\, \sin\alpha$ and $Q = m \,.\, \cos\alpha$. Consequently the expression (2.32) may be written

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} Q & P \\ -P & Q \end{pmatrix} . \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x'' \\ y'' \end{pmatrix} \tag{2.33}$$

The *inverse transformation* is that of determining the (x, y) coordinates whose (x′, y′) coordinates are already known. It may be required in converting from one map projection to another, because this is often a two-way process, as shown in Chapter 9. It can be shown that the inverse transformation corresponding to (2.33) is

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} Q' & -P' \\ P' & Q' \end{pmatrix} . \left(\begin{pmatrix} x' \\ y' \end{pmatrix} - \begin{pmatrix} x'' \\ y'' \end{pmatrix} \right) \tag{2.34}$$

where $Q' = \cos\alpha/m$ and $P' = \sin\alpha/m$.

Affine transformation

The assumption which is made in the Helmert transformation is that the scalar, m, is a single, unique value. In other words the ratio ab/AB is the

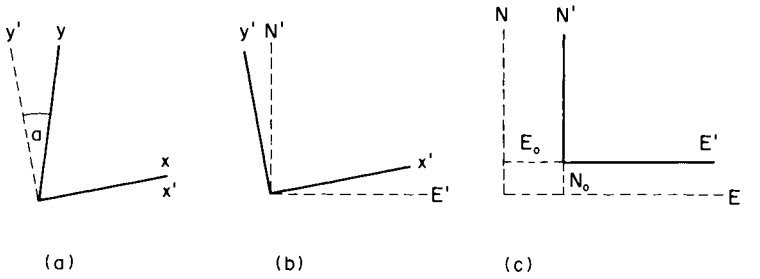


FIG. 2.10 The geometry of the affine transformation. Transformation (a) (x, y) to (x′, y′). Transformation (b) (x′, y′) to (E′, N′). Transformation (c) (E′, N′) to (E, N). (Source: Sprinsky, 1987.)

same whatever the directions of these lines. This is a reasonable assumption to make in some work, but it may not be justified for other jobs. For example in photogrammetry the location of image points on a film may be affected by deformation of the film base by stretching and shrinking, and this is not usually the same in all directions. In the extraction of positional information by digitising a paper map, the influence of differential stretching or shrinking of the paper must be considered. This may be large and unpredictable, as described by Maling (1989). For these applications it is desirable to use the affine transformation because this allows for different scales in the directions of the two axes, m_x and m_y. This may also be combined with small departures of the coordinate axes from the perpendicular, as illustrated in Fig. 2.10. Here we see that the (x, y) axes intersect at an angle $\gamma \neq 90°$. We need to determine six coefficients to solve equation (2.10).

Grid-on-Grid Calculations

The linear conformal transformation from one cartesian system to another is, as already stated, commonly used in cartography. From the nature of the first problem, all these transformations may be called *Grid-on-Grid Calculations*.

Although equation (2.33), with appropriate changes in notation from x to E and y to N, specify the final equations need to convert from the known (E′, N′) coordinates into the required (E, N) values, it is still necessary to determine suitable numerical values for P, Q, E″ and N″.

Provided that there are at least two points which are common to both systems, these terms can be calculated and used to convert as many additional points as are required. The method of solving the unknowns

may be carried out as below:

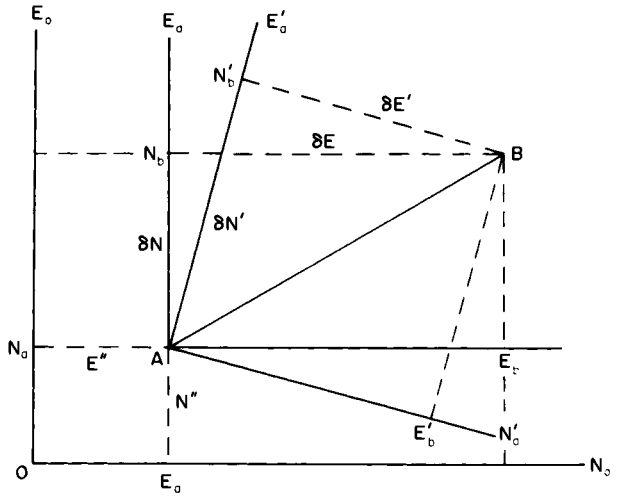


FIG. 2.11 The Grid-on-Grid problem. Stage 1, defining the relationship of two points *A* and *B*, whose coordinates on both grids are already known. E′ and N′ denote the initial grid; E and N denote the second grid to which other points are to be transformed.

In Fig. 2.11 the two points *A* and *B* are common to both grids. We use the following notation to described each point:

Point	1st grid	2nd grid
A	E'_a, N'_a	E_a, N_a
B	E'_b, N'_b	E_b, N_b

The coordinate differences between the two points may be expressed as follows, using the convention that the Greek letter δ signifies the difference between two coordinate values.

1st grid	2nd grid
$E'_a - E'_b = \delta E'$	$E_a - E_b = \delta E$
$N'_a - N'_b = \delta N'$	$N_a - N_b = \delta N$

These terms have the geometrical significance which is illustrated in Fig. 2.11. Using arguments similar to those already used to determine the effects of rotation and scale change upon the coordinates, it can be shown that

$$Q = [\delta E \,.\, \delta N' - \delta N \,.\, \delta E']/[\delta E'^2 + \delta N'^2] \tag{2.35}$$

$$P = [\delta N \,.\, \delta N' + \delta E \,.\, \delta E']/[\delta E'^2 + \delta N'^2] \tag{2.36}$$

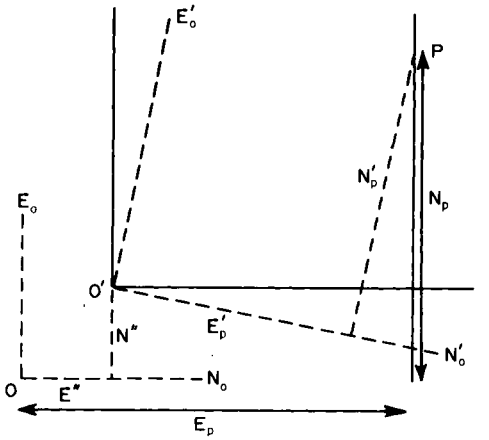


FIG. 2.12 The Grid-on-Grid problem. Stage 2, indicating the relationship of any point, *P*, whose coordinates on the initial grid (E'_P, N'_P) are known, to the second grid upon which it must be plotted.

The translation terms E″, N″, corresponding to x″ and y″ in equations (2.33) etc. may be found from

$$E'' = E_a - P.E'_a - Q.N'_a \tag{2.37}$$

$$= E_b - P.E'_b - Q.N'_b \tag{2.38}$$

$$N'' = N_a + Q.E'_a - P.N'_a \tag{2.39}$$

$$= N_b + Q.E'_b - P.N'_b \tag{2.40}$$

Hence the required equations to transform the (E′, N′) coordinates of any other point, *P* (Fig. 2.12) to the (E, N) system are

$$E = Q.E' + P.N' + E'' \tag{2.41}$$

$$N = -Q.E' + P.N' + N'' \tag{2.42}$$

which, converted into matrix notion provides an expression like (2.33).

The equations (2.35)–(2.38) have been given here without proof, but their derivation can be found, for example, in Ministry of Defence (1978). In Admiralty (1965) there is also described the method of solving the coefficients when there are three points common to both systems. If there are more than three common points, such as occurs in vector digitising and in the adjustment of aerial triangulation to many ground control points, the determination of the coefficients from only two or three of them is inadequate because the coordinates of any of those points may contain small errors and the use of them will introduce error into the transformation of all other points. Under these circumstances all of the data which are available for the determination of P and Q ought to be

taken into consideration. This involves a solution of the coefficients by the methods of least squares, which is a more sophisticated numerical solution based upon statistical error analysis.

The best procedure is to translate the axes of both system to a common origin at the centroid of the n points, obtained simply by determining the mean value of each coordinate. Thus for n points, labelled i = 1 . . . n,

$$E_G = \sum E_i/n \tag{2.43}$$

$$N_G = \sum N_i/n \tag{2.44}$$

with similar determinations for E'_G and N'_G.

The individual coordinates, E_i, N_i, E'_i, N'_i are now referred to these centroids as origin and the analysis of the most probable values for P and Q derived by standard routines. Modern textbooks on survey adjustments and computations, e.g. Hirvonen (1971), Cooper (1974), Mikhail (1976), Mikhail and Gracie (1981), and Methley (1986) all deal with the subject, and this book deals later (Chapter 19) with *polynomial transformations*, of which these are elementary examples.

The reader who is particularly concerned with the adjustment of vector digitised coordinates measured from paper maps which may also have been folded is referred specifically to the important paper by Sprinsky (1987).

CHAPTER 3

Coordinate reference systems on the sphere

'What's the good of Mercator's North Poles and Equators,
Tropics, Zones and Meridian lines?'
So the Bellman would cry: and the crew would reply
'They are merely conventional signs.'

Lewis Carroll, *The Hunting of the Snark*

Introduction

It has been assumed in Chapters 1 and 2 that the reader already knows something about the terms which are used to describe planes, arcs and angles on the earth. For example, the idea of latitude and longitude; parallels and meridians and the convergence of the meridians have been introduced without formal definition. However, it is desirable to consider these definitions and develop further our knowledge about the geometry of the earth. There are two reasons for this. First, we need to introduce a standardised system of algebraic notation for the different quantities which will be used throughout this book. Secondly it is necessary to demonstrate certain important geometrical differences between the sphere and the spheroid. In order to appreciate the distinctions to be made between these bodies it is essential to know precisely what is represented by planes, arcs and angles on each of them.

Some of the properties of a sphere have already been described in Chapter 1. These may be summarised as a preliminary to further definitions:

- A sphere is a solid body whose curved surface is everywhere equidistant from its centre.
- It follows that any sphere has constant radius.
- If a tangent plane meets any point on the curved surface, a line normal to this plane at the point of tangency is a radius to the centre of the sphere.
- The distance between two points on the sphere can be defined and measured either as the angular distance or the arc distance. There is

a simple relationship between the two measures of distance, which has been given in equation (1.02).

Definitions of planes, arcs and angles on the sphere

If a plane intersects a sphere, the resulting section of the curved surface which is traced on the plane is a circle. Two kinds of circle may be distinguished; a *great circle* and a *small circle*. If the intersecting plane passes through the centre of the sphere, the resulting section is the circle whose radius is the largest which can occur and is equal to the radius of the sphere itself. This is a great circle, illustrated by the outline of the sphere in Figs 1.02, 1.03, and many other later diagrams. Only one great circle can be drawn through any two points on the spherical surface which are not diametrically opposite to one another. The shorter arc of the great circle through two points is *the shortest distance between the points on the spherical surface*.

If the plane does not pass through the centre of the sphere, the radius of the resulting circle is less than that of the sphere. This is a small circle, shown in Fig. 3.01 by the line *EFGH*. These points all lie on the circumference of a circle with centre *O'*.

The *axis* of any circle is the straight line passing through the centre of the sphere at right angles to the plane of the circle. Thus, in Fig. 3.01, the line *POP'* is the axis to the great circle *DABC*. From the definition that only one great circle can be drawn through a pair of points that are not diametrically opposite, it follows that the axes of two or more great circles cannot coincide. However, one great circle and any number of small circles can have a common axis. From the definition of an axis it follows that, in this special case, the planes of the great circle and all the small

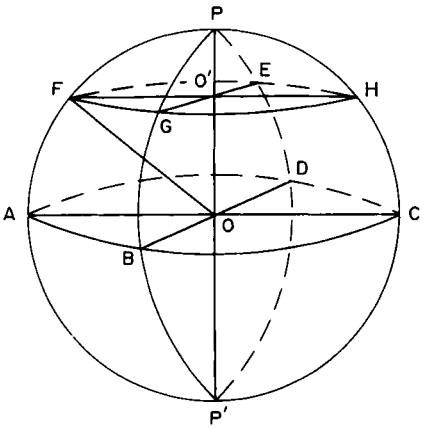


FIG. 3.01 Great circles and small circles on the sphere.

circles will be parallel to one another. Moreover if the planes are parallel, the circumferences of the circles are also parallel.

The *poles* of any circle are the points where the axis to a circle intersects the surface of the sphere. These are shown in Fig. 3.01 by the points P and P', which are the poles to the great circle $DABC$. From the definitions that a sphere has constant radius and that the section of a great circle passes through the centre of the sphere, it follows that the poles to a great circle are equidistant from its plane. From the corresponding definition of a small circle, clearly one pole is nearer than the other. If the great circle $DABC$ is further described as a *primary* or *primitive great circle*, then any great circle which passes through its poles may be called a *secondary great circle*. Since the poles are diametrically opposite to one another any number of secondaries may be specified. In Fig. 3.01 the great circle arcs $PFAP'CH$ and $PGBP'DE$ are both secondaries to the great circle $DABC$. Since the axis to the primary great circle coincides with the plane of each secondary, it follows that the plane and therefore the circumference of the primary great circle will also have planes and circumferences which are perpendicular to the secondaries to that great circle.

Geographical coordinates

Since the earth is a rotating body, the obvious datum from which we may define its geometry is its axis of rotation. This axis intersects the surface at two points which are the poles to a primary great circle whose plane is perpendicular to the axis. The primary great circle is the *equator* and its poles are the *north* and *south geographical poles*. The secondaries to the equator are not given a single name but the word *meridian* describes each semicircle of a pair which together form a single secondary. The word meridian should be used in the restricted sense of being the arc of any great circle passing through and limited by the geographical poles. The complete secondary comprises one meridian together with its *anti-meridian*.

It follows from the use of angles at the centre of a sphere to measure distances between points on the curved surface, that a system of three-dimensional polar coordinates may be used as a method of locating position with respect to the centre of the sphere as origin. By extension of the concept of plane polar coordinates described in Chapter 2, a point may be located in space if we know *two* vectorial angles and the radius vector. These are known as *spherical polar coordinates* in mathematics. However, all points on the surface of the sphere are equidistant from the centre. Therefore the radius vector is always equal to the radius of the sphere and serves no useful purpose, in this special case of coordinate location. Thus coordinate position on the spherical surface is uniquely

defined by means of two vectorial angles. For these, two orthogonal planes are chosen which intersect at the origin (i.e. the centre of the sphere). One plane has already been defined and is the plane of the equator. This is used as the datum of measurement of the vectorial angle which we know as *latitude*. The other plane is that of the meridian chosen as zero *longitude*.

Latitude

The latitude of a point may be formally defined as *the angle measured at the centre of the earth between the plane of the equator and the radius drawn to the point*. It is, for example, the angle AOQ in Fig. 3.02. This definition applies only to latitude measured in a true sphere. It will be seen later that *it is necessary to use different definitions for latitude on the spheroid.* For most practical purposes, latitudes may be expressed in sexagesimal units north and south of the equator. Centisimal units are used for this purpose in certain countries or for certain purposes, but it is important to realise that, just because a nation had adopted the metric system and decimal notation for most other kinds of measurement, this does not automatically mean that angles are measured in grads. Geographical coordinates expressed in centisimal units are the exception rather than the rule. Algebraically the angle is usually denoted by φ, and this symbol is used to mean latitude throughout the present book. In order to use a logical sign convention for algebraic purposes it is customary to regard north latitude as $+\varphi$ whereas south latitude is $-\varphi$.

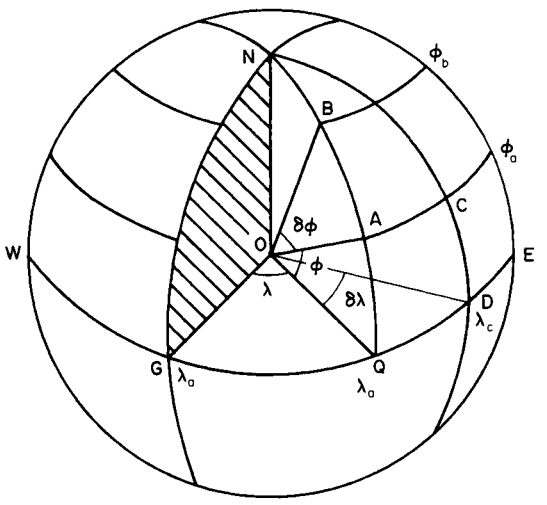


FIG. 3.02 Latitude and longitude on the sphere. The plane of the Greenwich Meridian is shaded.

The *difference in latitude* between any two points is the quantity

$$\delta\varphi = \varphi_a - \varphi_b \tag{3.01}$$

where the symbol δ indicates a finite angular difference between the latitude of some point $A = \varphi_a$ and another point $B = \varphi_b$, both angles measured from the plane of the equator according to the definition given above. If we need to refer to a very (infinitely) small change in latitude, we introduce the notation of the calculus and state that as $\delta\varphi \to 0$ (which is the mathematical shorthand for the statement 'as the difference in latitude approaches zero') it may be represented by $d\varphi$.

For any given value of φ there are an infinity of points on the surface of the earth each of which makes this angle with the plane of the equator. The locus of these points is the circumference of a circle, the plane of which is parallel to that of the equator. Consequently it may be called a *parallel of latitude*, or simply a *parallel*. It follows that as the plane of this circle is parallel to the equatorial plane they share a common axis. Because the equator is a great circle, it follows that any parallel of latitude other than the equator must be a small circle.

Since the plane of the equator is perpendicular to the earth's axis of rotation, the angle measured at the centre of the sphere between this axis and the radius to a point in latitude φ, such as NOA in Fig. 3.02, is the complement of the latitude ($90^\circ - \varphi^\circ$ or $\pi/2 - \varphi$ radians). This angle is therefore called the *colatitude* of the point and will be denoted algebraically by χ.

Longitude

The longitude of any point on the earth's surface represents the second vectorial angle required to define position. This may be defined as *the angle measured in the plane of the equator between the plane of the meridian through the point and the plane of some other meridian selected as datum.*

The choice of a datum meridian for measurement of longitude is arbitrary. Although we are generally accustomed to the use of the meridian passing through the former site of the Royal Observatory at Greenwich as the *Prime Meridian* for measurement of longitude, any other meridian would be equally satisfactory. From the point of view of a national survey and the production of topographical map series it can be argued that no particular national advantage is served by relating longitude to the Greenwich Meridian. For example, the longitude of Paris is used as the datum for French maps, the meridians of Oslo, Rome and Leningrad (Pulkova Observatory) have been respectively employed for the origin of longitude on maps of Norway, Italy and the USSR. Sometimes a more or less arbitrary origin has been used. The classic example of this was the

use of the approximate meridian of Ferro in the Canary Islands as the datum for longitude, first in France, later in the Austro-Hungarian and German Empires and therefore to quite modern maps of Austria, Czechoslovakia, and Hungary. To confuse the issue further, precise definition of the longitude of Ferro varied according to how a particular survey organisation originally interpreted it.

For other purposes, particularly in navigation and astronomy, where the apparent movements of heavenly bodies with time must be referred to longitude, it is extremely inconvenient to have more than one origin for measurement. The use of the Greenwich Meridian as the Prime Meridian was agreed internationally in 1884, and this remains the preferred datum.

Longitude is measured from this plane, normally in sexagesimal units east and west of Greenwich. The algebraic symbol used for the angle is λ. The sign convention is that $+\lambda$ indicates east longitude whereas $-\lambda$ means west longitude.

The term

$$\delta\lambda = \lambda_a - \lambda_c \tag{3.02}$$

signifies the *difference in longitude* between two places, $A = \lambda_a$ and $C = \lambda_c$. This is the angle DOQ in Fig. 3.02. The symbol δ again indicates a finite difference in longitude and for an infinitely small increment in longitude we use $d\lambda$. Frequently in the derivation of general expressions for map projections it is convenient to refer longitude to some meridian other than Greenwich such as the central meridian of a map. Then we denote this meridian as λ_0.

Graticule

The resulting network of parallels and meridians which comprise the system of geographical coordinates is known as a *graticule* or net, but with reference to the earth's surface and to the representation of it on a plane surface by means of a map projection. A *graticule intersection* is the point where the parallel φ intersects the meridian λ, and is referred to by its geographical coordinates (φ, λ). The convention of describing these coordinates in the order latitude-followed-by-longitude is universally accepted.

Position in geographical coordinates is by far the best-known method of providing unique reference of location in geography, navigation and all the other sciences and technologies which are concerned with the earth. The network of parallels and meridians on the map or chart constitutes a form of geometrical control to map use which is understood by most cultures and at many different levels of education. It is a reference system taught to schoolchildren early in their geographical education. It

follows that the network of parallels and meridians as remembered from maps in a school atlas and from a globe ought to remain an important spatial frame of reference for map use in later life. Moreover the graticule has historical importance, for it is much older than the general concept of spherical polar coordinates or other systems. Some kind of representation of the parallels, in the form of zones, has been used since the time of Marinus and Ptolemy in the first century AD. Few worthwhile maps have been produced since the early seventeenth century which do not show some kind of graticule.

Nevertheless it would be wrong to suppose that geographical coordinates are the only method of defining position upon the earth's surface. Reference has already been made to the generalised system of polar coordinates in three dimensions, of which the (φ, λ) system is a special case. Another system of spherical polar coordinates which are, again, suitably simplified for representation of the curved surface of a sphere are the *bearing and distance coordinates*, which have particular value in the construction of map projections. These are studied in Chapter 7. A third and quite different system of location is by *three-dimensional cartesian coordinates* which differ from plane rectangular coordinates by the addition of a third, Z, axis which is perpendicular to the other two and, for both sphere and spheroid, corresponds to the axis of rotation. This is described in Chapter 4.

Angles on the sphere

Having established the properties of geographical coordinates as the primary method of location, it is now desirable to introduce some additional concepts about the geometry of the sphere.

A *spherical angle* is the inclination, at their point of intersection, of two arcs of great circles measured on the curved surface of the sphere. It is also equal to the plane angle formed between two tangents, drawn at the point of intersection, one to each great circle. Thus in Fig. 3.03, the spherical angle between the two great circles PA and PD is the angle DPA, which is equal to the plane angle KPJ. For the purpose of the present study spherical angles are encountered in two forms. The first is to permit an alternative definition of longitude. That given on page 51 describing longitude as an angle measured at the centre of the sphere in the plane of the equator is so worded to emphasise that geographical coordinates are a form of polar coordinates and the vectorial angles should therefore be measured at the origin of the system. However, we can see from Figs 3.01, 3.02 and 3.03 that the angle λ can be measured anywhere on the earth's axis of rotation, provided that this angle is measured in a plane parallel to the equator. Thus the angle $FO'G = AOB$ in Fig. 3.01 and the longitude can be measured in the plane of any parallel

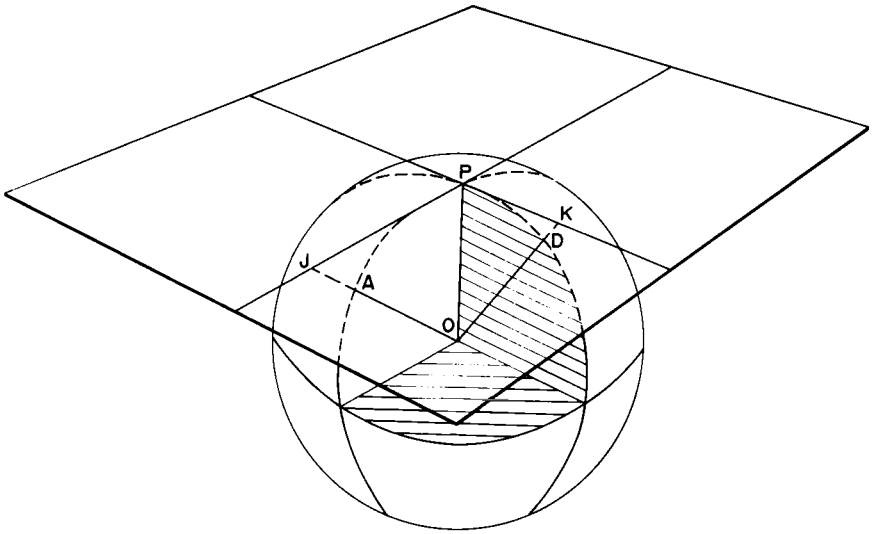


FIG. 3.03 Definition of a spherical angle.

of latitude. Extending this argument to the geographical poles, such as *P* in Figs 3.01 and 3.03, it follows that the plane through *P* which is parallel to the equator is also the tangent plane at *P*. Hence longitude can be measured as the plane angle *KPJ* in Fig. 3.03 or as the spherical angle *APD*.

The second important kind of angle encountered on the earth's surface is the *azimuth* or *bearing* of one point measured from another. This introduces the concept of *direction* on the earth and also some rather fine distinctions of definition. Consider the three points *N*, *A* and *B* illustrated in Fig. 3.04. The point *N* is the North Pole, so that the great circle arc *NA* represents part of the meridian through *A*. Similarly the arc *NB* is part of the meridian through *B*. The line *AB* represents the shortest distance between *A* and *BV* and is therefore the arc of the great circle. Hence the *spherical triangle* has been formed by the intersection of three great circle arcs.

Azimuth and bearing

Azimuth may be defined as: * *the spherical angle between any great circle and a meridian.* Thus the angle *NAB* represents the azimuth of *B* measured at *A*; the angle *NBA* represents the azimuth of *A* from *B*. In the southern hemisphere the equivalent azimuths are *SAB* and *SBA*. We have seen in Chapter 2 that in navigation, surveying and cartography the usual convention is to measure angles according to the 360° or 400^g circle in a

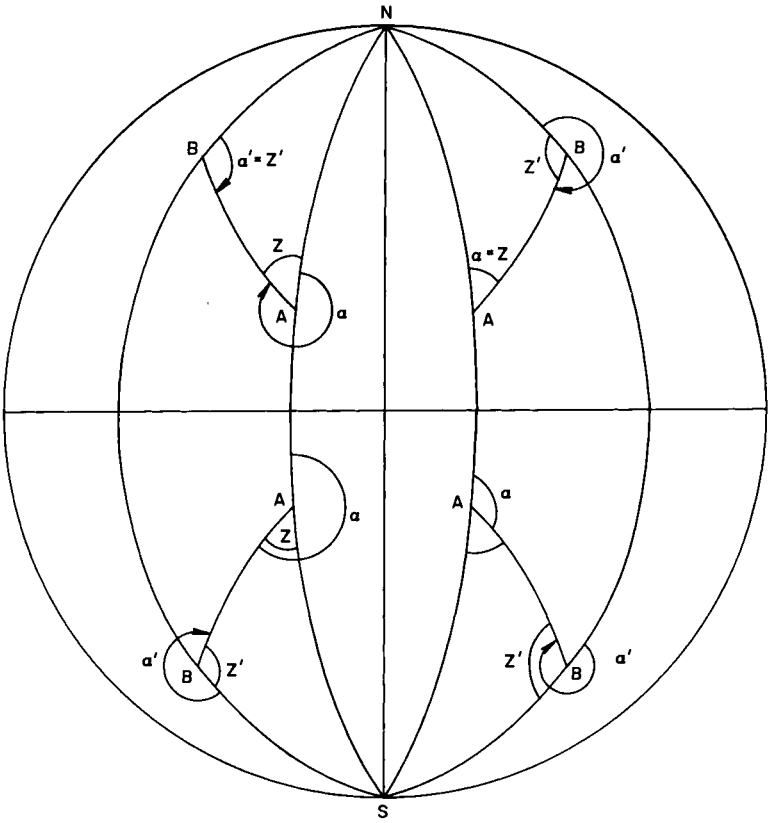


FIG. 3.04 The definition of bearing (α), reverse bearing (α'), azimuth (Z) and reverse azimuth (Z') for four different versions of the arc AB on the spherical surface. The description in the text specifically refers to the arc AB in the north-east part of the diagram.

clockwise direction. This convention is contained in the formal definition of *bearing* which is: * *the horizontal angle at a given point measured clockwise from a specific reference point to a third point.* If the specific reference point is the North Pole, then we have the definition of a *true bearing* which is: * *the direction of an object from a point, expressed as a horizontal angle measured clockwise from true North.* In the north-east quadrant of Fig. 3.04 the azimuth at A is the acute angle Z, measured clockwise from the meridian AN. Here $\alpha = Z$ and the angle also represents the true bearing of B from A.

From the definition of azimuth, the angle NBA represents the azimuth of A from B. However, according to the clockwise convention of bearing, the true bearing of A from B is the clockwise angle at that point indicated as α'. In the southern hemisphere of Fig. 3.04 the azimuths would have

been referred to the South Pole, whereas the true bearings are still measured clockwise from true north. Such distinctions are not always made in the literature.

Spherical triangle

A spherical triangle is the figure formed by the intersection of any three arcs of great circles. Like a plane triangle it comprises six parts: three angles and three sides. The notation which is used to describe these parts for simple algebraic expression is the same as plane geometry. Thus, in Fig. 3.05 we may write for the angles $ABC = B$, $ACB = C$ and $BAC = A$. Similarly the three sides are described as $CB = a$, $AC = b$ and $AB = c$. Since the arc of a great circle has length proportional to the radius of the sphere, and since all the sides belong to the same sphere, it is sufficient to define the lengths of the sides only by angular distance. This, as we saw in Chapter 1, is measured at the centre of the sphere by the angles between the radii drawn to the three points.

Many of the fundamental properties of a spherical triangle are equivalent to those of a plane triangle. To quote just one example, the greatest side of a spherical triangle is always opposite the largest angle. However, the properties of a spherical triangle differ from those of a plane triangle in one extremely important respect. *The sum of the three angles of a spherical triangle does not equal 180° but is always greater.* The difference between the sum of the angles and 180° is known as the *spherical excess*, and is proportional to the area of the triangle. For example, the spherical triangle representing 1/8 of the total surface area of the sphere, in which all the sides and angles are equal to 90°, has spherical excess amounting to 90°. The existence of spherical excess profoundly influences the

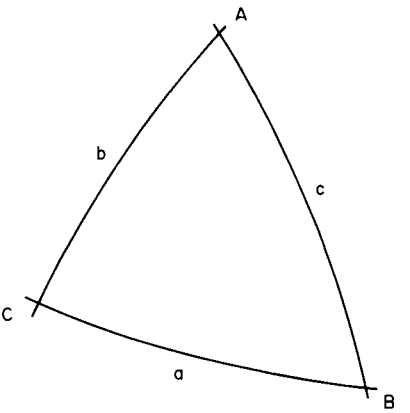


FIG. 3.05 The spherical triangle.

methods of spherical geometry and trigonometry. *It is not possible, as in plane geometry, to determine the value of an unknown angle of a spherical triangle by subtracting the sum of two known angles from 180°.*

Determination of the unknown parts of a spherical triangle

Just as plane trigonometry can be used to determine the length of an unknown side, or the size of an unknown angle in a plane triangle, so equivalent calculations can be used to solve unknown parts of spherical triangles. The methods of solution may be grouped under the heading *spherical trigonometry*. It is a branch of mathematics which is particularly important in certain practical applications such as navigation, surveying and astronomy. For example, position finding by astronomical methods is almost wholly dependent upon the solution of spherical triangles on the celestrial sphere and the earth. The subject is also important to the study of map projections. Space does not permit full derivation of the formulae which are useful to the cartographer. We therefore refer to certain other works such as Admiralty (1960), Clough-Smith (1966) and Cotter (1969) which are devoted to spherical trigonometry or navigation.

The two most important formulae of spherical trigonometry, from which all others may be derived, are:

- The Cosine or Fundamental Formula;
- The Sine Formula.

Cosine formula

This gives the relationship between one unknown side of a spherical triangle when the other two sides and their included angle are known. For the triangle ABC illustrated in Fig. 3.05 this may be written for three possibilities:

1. Unknown side a; known sides b and c; known angle A

$$\cos a = \cos b \,.\, \cos c + \sin b \,.\, \sin c \,.\, \cos A \qquad (3.03)$$

2. Unknown side b; known sides a and c; known angle B

$$\cos b = \cos a \,.\, \cos c + \sin a \,.\, \sin c \,.\, \cos B \qquad (3.04)$$

3. Unknown side c; known sides a and b; known angle C

$$\cos c = \cos a \,.\, \cos b + \sin c \,.\, \sin b \,.\, \cos C \qquad (3.05)$$

On the other hand, if the three sides of the triangle are known, the formulae may be modified to solve one unknown angle,

$$\cos A = [\cos a - \cos b \,.\, \cos c]/[\sin b \,.\, \sin c] \qquad (3.06)$$

These formulae give a single, unambiguous, result for the unknown side or angle. By convention, the sides and angles of a spherical triangle cannot exceed 180°. Therefore the result must be the cosine of an angle in the first or second quadrant. If the answer is positive, this indicates that the angle lies in the first quadrant ($0° < \alpha < 90°$) but if it is negative this means that the angle lies in the second quadrant ($90° < \alpha < 180°$). These sign differences are important in computing; a subject to which we shall return later.

Sine formula

This has the form

$$\sin a/\sin A = \sin b/\sin B = \sin c/\sin C \tag{3.07}$$

Thus, knowing three parts (sides and angles) for any pair of ratios, it is possible to find the unknown part. For example, if a, b and B are known

$$\sin A = (\sin a \,.\, \sin B)/\sin b \tag{3.08}$$

$$= \sin a \,.\, \sin B \,.\, \operatorname{cosec} b \tag{3.09}$$

The sine formula suffers from the important disadvantage that there is ambiguity about the part found, for $\sin A = \sin (180° - A)$. Various rules are given, in the textbooks of spherical trigonometry, which attempt to overcome this difficulty.

The lengths of arcs on the earth's surface

There are three kinds of arc measurement which are important to the study of map projections. These are:

- the length of the arc of a meridian;
- the length of the arc of a parallel;
- the length of the arc of any great circle.

The first two are essential to the derivation of the scale errors and distortions in the directions of the meridian and parallels at a point. Knowledge about these is an essential prerequisite to the derivation of any map projection which is intended to satisfy one of the mathematical properties described in Chapter 4. The third kind of measurement is more commonly thought of as a procedure in navigation and other kinds of qualitative map or chart use. This is the way to determine the great circle distance between two places when a high order of accuracy is not required and the spherical assumption suffices. However, this general expression for determining the arc of any great circle arises in the transformation from geographical into bearing and distance coordinates, as described in Chapter 9 (pp. 178–183).

The length of the arc of a meridian

This problem was mentioned superficially in Chapter 1 to indicate the methods of astro-geodetic arc measurement as a means of determining the Figure of the Earth. From equation (1.02), and using the algebraic notation introduced in this chapter, various meridional arc relationships may be expressed as follows (Fig. 3.06):

- The length of the arc measured from the plane of the equator to the point F in latitude φ_f:

$$s_m = R \,.\, \varphi_a \tag{3.10}$$

- The length of the arc measured from the nearer pole to the same point:

$$s'_m = R \,.\, \chi_f \tag{3.11}$$

- The arc distance between two points, $A = (\varphi_0, \lambda_a)$ and $F = (\varphi_f, \lambda_a)$ both of which lie on the same meridian.

$$s''_m = R \,.\, \delta\varphi \tag{3.12}$$

where $\delta\varphi = \varphi_0 - \varphi_f$.

Following the derivation of (1.02) all the angles in equations (3.10)–(3.12) are expressed in radians.

The length of the arc of a parallel

It has been shown that a parallel of latitude is a small circle. This has radius r and, by definition $r < R$. Thus, for any given angular distance, the arc distance along a parallel is less than the corresponding arc distance along the equator. In Fig. 3.06, for example, NFA represents the meridian λ_a and NGB is the meridian λ_b. Therefore the angle $AOB = FO'G = \delta\lambda$. From equation (1.02):

$$AB = R \,.\, \delta\lambda \tag{3.13}$$

and

$$FG = r \,.\, \delta\lambda \tag{3.14}$$

In the right-angled triangle OFO', $OF = R$ and $O'F = r$. Moreover the angle $O'OF$ is the colatitude, χ of F. Therefore

$$r = R \,.\, \sin\chi \tag{3.15}$$

$$= R \,.\, \cos\varphi \tag{3.16}$$

Consequently the arc distance along the parallel of latitude φ is

$$s_p = R \,.\, \cos\varphi \,.\, \delta\lambda \tag{3.17}$$

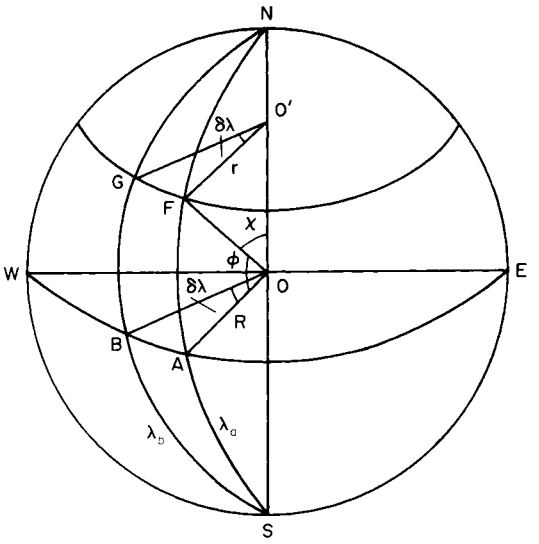


FIG. 3.06 The relationship of the radius of a parallel, r, to the radius of the sphere, R.

For an arc on the equator, we put $\varphi = 0°$ so that $\cos \varphi = 1{\cdot}0$. Then (3.17) becomes

$$s_e = R \,.\, \delta\lambda \tag{3.18}$$

which is the result we would expect from the definition of the equator as a great circle.

The length of the arc of any great circle

In equation (1.02) we used z to indicate the angular distance between two points. We now return to the general case of the length of any great circle arc and we use this letter to indicate the unknown angular distance between two points which lie in different latitudes and longitudes. Thus, if $A = (\varphi_a, \lambda_a)$ and $B = (\varphi_b, \lambda_b)$, as illustrated in Fig. 3.04, we have to solve the spherical triangle NAB to find the unknown side $AB = z$.

The two known sides of the triangle are the meridional arcs NA and NB, which are of length χ_a and χ_b respectively. The spherical angle $ANB = \lambda_a - \lambda_b = \delta\lambda$ is also known. From the cosine formula (3.03)–(3.05),

$$\cos z = \cos \chi_a \,.\, \cos \chi_b + \sin \chi_a \,.\, \sin \chi_b \,.\, \cos \delta\lambda \tag{3.19}$$

This is more conveniently expressed in terms of latitude rather than

colatitude. Thus

$$\cos z = \sin \varphi_a . \sin \varphi_b + \cos \varphi_a . \cos \varphi_b . \cos \delta\lambda \qquad (3.20)$$

and, finally,

$$s = R . z \qquad (3.21)$$

Conversion of arc length into linear distance

In order to convert any of the values of s, s_e, s_p or s_m into linear units, we require a suitable value for R, which must be determined from the radii of the adopted Figure of the Earth. We shall see later that there are many ways of obtaining a suitable radius, but, in order to appreciate the significance of the different measures, it is necessary to know more about the geometry of the spheroid. Therefore we do not compare the different methods or their results until the end of Chapter 4, where they are listed in Table 4.02, page 79.

It will be seen that for a given Figure of the Earth (the International Spheroid in Table 4.02) there are substantial differences between the results. Before attempting to make a choice it must also be appreciated that we have made the initial assumption that the earth is a perfect sphere. From the point of view of constructing a map to a specified scale, this assumption naturally influences all subsequent calculations so that use of R correct to the nearest metre, as in Table 4.02, may introduce a spurious appearance of accuracy to some calculations. In the example of making calculations to construct the graticule, the reduction of the metric values by a scale fraction which may be less than 1/1 000 000, will result in any small niceties in the metric values for R being wholly absorbed in the plotting process. Then it is sufficient to use either of the most commonly used values for R. These are the *authalic radius*, being the radius of a sphere having the same surface area as the chosen Figure of the Earth, and the radius of the sphere having the same volume as the chosen figure for a sphere based upon these two determinations made from the International Spheroid are 6371228 m and 6371221 m, respectively. *For most practical applications in small-scale cartography it is sufficient to take the radius of the sphere as being 6371·2 km.* Without prejudice to these comments it is also necessary to appreciate that in some geodetic applications, including the projection of the spheroid to a plane map, we sometimes employ an *auxiliary* sphere to make certain transformations. This is a part of the spherical surface which is considered to be tangential to some part of the spheroid. For these purposes, as we shall see in Chapter 16, for example, precise definition of R is essential.

In some theoretical work with map projections it is not necessary to convert angular distances into their linear equivalents. It is therefore

sufficient to derive all projections in terms of sphere of unit radius (R = 1) and then convert the numerical values obtained by a factor which corresponds to the radius of the earth in millimetres reduced to the required map scale. The method is described in detail in Chapter 8.

Angles on the earth's surface

Determination of azimuth

From the definition of azimuth given on page 54, this is the angle $NAB = Z$ in the simplest case of the north-east quadrant. The value of Z may be determined from a modified version of the cosine formula (3.06). Using the same notation employed in (3.19)

$$\cos Z = [\cos \chi_b - \cos \chi_a \,.\, \cos z]/[\sin \chi_a \,.\, \sin z] \tag{3.22}$$

$$= [\sin \varphi_b - \sin \varphi_a \,.\, \cos z]/[\cos \varphi_a \,.\, \sin z] \tag{3.23}$$

Alternatively, from the sine formula (3.09)

$$\sin Z = \cos \varphi_b \,.\, \sin \delta\lambda \,.\, \operatorname{cosec} z \tag{3.24}$$

Both of these equations contain terms in z. If z is not required, then the preliminary calculation of it can be avoided. It is possible to combine equations (3.19) and (3.22) which, after some algebraic manipulation, results in the equation

$$\cot Z = \cos \varphi_a \,.\, \tan \varphi_b \operatorname{cosec} \delta\lambda - \sin \varphi_a \,.\, \cot \delta\lambda \tag{3.25}$$

which is independent of z.

Convergence of meridians

It should be noted that the bearing from *B* to *A*, denoted by the clockwise angle *NBA*, is not the reciprocal of α. In other words $[180^\circ - \alpha] \neq NAB$, but differs by the angle γ, shown in Fig. 3.07. This leads to the interesting and important conclusion that the azimuth of any great circle which crosses the meridians obliquely can only be defined uniquely at the point where it is measured. In other words, *the bearing of a great circle arc changes continuously*. The reason for this is the *convergence of the meridians*. On the equator the arc distance between two meridians is, as we have seen in equation (3.18), s_e. At the geographical poles the corresponding arc distance is zero. On the equator, two meridians λ_a and λ_b are perpendicular to it. At the poles the same meridians intersect to make the spherical angle $\delta\lambda$. The angle of convergence (or *convergency*) between the meridians in any intermediate latitude may be expressed by the angle γ. The value of γ varies with latitude and it can be shown that

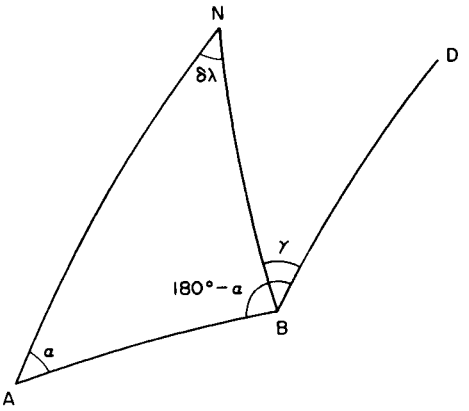


FIG. 3.07 The relationship between bearing (α), reverse bearing ($180° - \alpha$) and convergence of the meridians (γ) on the sphere.

it varies according to

$$\gamma = \delta\lambda \,.\, \sin \varphi \tag{3.26}$$

For any line lying between the parallels φ_a and φ_b it is usual to express the convergency of their meridians in terms of the *mean latitude* as

$$\gamma = \delta\lambda \,.\, \sin [(\varphi_a + \varphi_b)/2] \tag{3.27}$$

This formula is adequate for most purposes in navigation but it is too crude for use in surveying. More precise versions are given in equations III.13 and III.39 in Appendix III, pp. 445 and 447.

CHAPTER 4

The geometry of the spheroid

> Therefore if *APBQ* represent the figure of the earth, now no longer spherical, but generated by the rotation of an ellipse about its lesser axis *PQ*; and *ACQqca* a canal full of water, reaching from the pole *Qq* to the centre *Cc* and thence rising to the Equator *Aa*; the weight of the water in the leg of the canal *ACca* will be the weight of water in the other leg *QCcq* as 289 to 288, because of the centrifugal force arising from the circular motion sustained and takes off one of the 289 parts of the weight (in the one leg), and the weight of 288 in the other sustains the rest....
>
> Isaac Newton, *Principia Mathematica*, 1687

Introduction

We now consider the definition and expression of the planes, arcs and angles on the spheroid corresponding to those studied in Chapter 3 with respect to the sphere. We have already seen in Chapter 1 that an ellipsoid of rotation may be defined by the length of the major semi-axis, a, and the flattening, f. We may also use two other combinations:

- the lengths of the two semi-axes, a and b;
- the length of the semi-major axis, a and the eccentricity, e^2, to be defined below.

It follows from Fig. 1.01 that the meridional section of the figure is an ellipse but that the equator is represented by mean of a circle of radius a. With the exception of the equator and the parallels of latitude there are no circles defined by plane sections through the ellipsoid. The curve which corresponds to any great circle on the sphere may be called a *geodesic*, but there are mathematical difficulties in defining the word. We will avoid these difficulties in the elementary exposition of this chapter by referring only to an arc.

Spheroidal parameters

The circumference of an ellipse may be defined as the locus of points, the sum of whose distances from two fixed points is constant and equal to

2a. These two points are known as the foci of the ellipse. They lie on the major axis and are indicated by the points F_1 and F_2 in Fig. 3.08. The eccentricity is equal to the ratio of OF_1/OW. From the right-angled triangle F_1NO

$$OF_1 = a^2 - b^2 \tag{4.01}$$

Since $e = OF_1/OW$, it follows that the *first eccentricity* of the spheroid

$$e^2 = (a^2 - b^2)/a^2 \tag{4.02}$$

The numerical value of e^2 for the earth is about 0·0067..., but the more precise determination depends upon the values of a and b for the selected Figure of the Earth.

A number of other related parameters are also used in geodesy. These include

- *Second eccentricity*,

$$e'^2 = (a^2 - b^2)/b^2$$

- *Polar radius of curvature*,

$$c = a^2/b$$

- *n*

$$n = (a - b)/(a + b)$$

Obviously these parameters are closely related. We note the following simple algebraic relationships between e^2, e'^2, f, n and c:

$$e^2 = 2f - f^2 = 4n/(1+n)^2 \tag{4.03}$$

$$f = 1 - (1 - e^2)^{1/2} = 2n/(1+n) \tag{4.04}$$

$$n = f/(2 - f) = [1 - (1 - e^2)]^{1/2}/[1 + (1 - e^2)]^{1/2} \tag{4.05}$$

$$e'^2 = (2f - f^2)/(1 - f)^2 = e^2/(1 - e^2) = 4n/(1 - n^2) \tag{4.06}$$

$$a = c(1 - n)/(1 + n) \tag{4.07}$$

These may appear in the algebraic derivation of the projections of the spheroid and may also be used to simplify computation. A variety of other coordinate systems are now used in geodesy, some of which will be introduced later. A summary of many of the others may be found in a useful paper by Soler and Hothem (1988).

Latitude on the spheroid

We have already noted in Chapter 1 that the radii of curvature at any point on an ellipsoid must be normal to the tangent plane to the point.

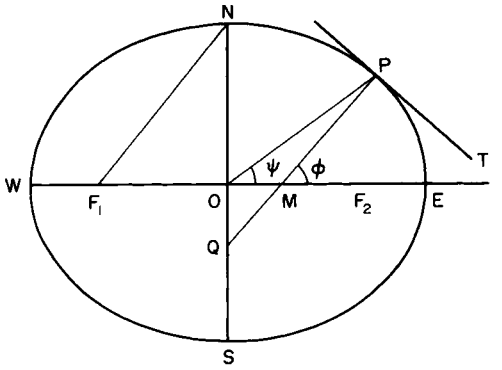


FIG. 4.01 The definition of latitude on the spheroid.

Thus, for a point P in Fig. 4.01, the normal to the tangent plane is the line PQ. This line intersects the major axis of the ellipse at M and therefore makes the angle PME with the plane of the equator. By contrast, the line PO drawn to the point of intersection of the two axes makes the angle POE with the major axis in the plane of the equator. Clearly these angles differ, but both of them correspond in part to the definition for latitude on the sphere. Thus the angle POE corresponds to the idea that it is the angle measured at the centre of the earth, but the angle PME corresponds to the idea that the angle is measured between the plane of the equator and the radius drawn to the point. We distinguish between the two different definitions for latitude as follows:

Geocentric latitude

Geocentric latitude is the angle, measured at the point of intersection of the axes of the spheroid, between the plane containing the major semi-axis and the straight line to some point on the surface of the spheroid. This is the angle POE which is usually denoted by ψ.

Geodetic latitude

Geodetic latitude is the angle between the major axis of the spheroid and the normal to the tangent plane at any point on the surface of the spheroid, measured at the point of intersection of the normal with the equatorial plane. This is the angle PME which is denoted by φ.

There is a relationship between these two angles which may be expressed in terms of eccentricity, but this is not of direct importance to us in the present context apart from observing that the difference between the two definitions varies with latitude and is greatest in latitude 45° where it amounts to nearly 12′ of arc.

Geodetic latitude is the more important quantity, and this is the variable which enters into most subsequent calculations relating to the spheroid. For most practical purposes, geographical coordinates on the spheroid are taken to be the (φ, λ) system, where φ is the geodetic latitude and λ is the longitude. The definition of longitude on the spheroid is the same as that for the sphere.

Auxiliary latitudes

In addition to geodetic and geocentric latitudes there are four further definitions of latitude to be considered. These are used to map the spheroid to an *auxiliary* sphere according to certain mathematical principles which we shall come to recognise later as being special properties of projections. Formulae for the spherical form of a given map projection may be adapted for use with the spheroid by substitution of one of the various auxiliary latitudes in place of geodetic latitude. Geocentric latitude is one of these; so, too, is the reduced latitude, u, defined in (4.33) as the starting point for calculating three-dimensional cartesian coordinates of points on the spheroidal surface. There are three more important possibilities which we should know about relating to the three *special properties* of map projections, which we shall encounter in Chapter 5, namely *conformality*, *equivalence* and *equidistance*.

- *Conformal latitude* is used to map the spheroid conformally upon an auxiliary sphere.
- *Authalic latitude* is used to map the spheroid to an auxiliary sphere in such a way that the sphere is equal in area to that of the spheroid.
- *Rectifying latitude* or *equidistant latitude* is used to map the spheroid upon an auxiliary sphere in such a way that correct distances along the meridians have been preserved.

These auxiliary latitudes were derived by Adams (1921) using series in geodetic latitude, φ and eccentricity e^2. The most recent summary of this work is to be found in Snyder (1987a). There is a small difference between each of these auxiliary latitudes and geodetic latitude, which is zero at the equator and the poles, reaching a maximum in latitude 45°. The size of this difference varies too with the adopted Figure of the Earth. Table 4.01, which is an extract from Adams's original work, indicates the maximum differences obtained from the Clarke 1866 spheroid.

The reader should note some inconsistency in the description of conformal latitude. Adams (1921) called this *isometric latitude*. However, from the time that Lee (1946) first drew attention to the inconsistency, we have used the term *orthomorphic* or *conformal latitude* to mean this auxiliary latitude, and retained the term isometric latitude for an entirely different purpose; as the parameter to transform Mercator's projection

TABLE 4.01 *Comparison of the difference between geodetic latitude, φ, and the five auxiliary latitudes of the spheroid for latitude 45° where the difference is maximum. The difference is auxiliary–geodetic.*

Geocentric	Reduced	Conformal	Authalic	Rectifying
−11′ 40″.5	−5′ 45″.0	−11′ 40″.0	−7′ 47″.0	−8′ 45″.3

from the spheroid to the sphere. For the most recent study of this parameter, see Bowring (1990a).

The radii of curvature of an ellipsoid

The concept of radius applied to the ellipsoid is more complicated than for the sphere. The first difficulty is that two radii of curvature may be defined at any point; the second is that both of these radii vary with latitude. The two radii at a point such as A are

- *Meridional radius of curvature.* This is the radius of curvature of the ellipse NAE at the point A. This quantity is usually referred to as ρ.
- *Transverse radius of curvature.* This is the radius of the curve formed by a plane intersecting the ellipsoid at A which is normal to the surface and also perpendicular to the meridian at the point. This is a difficult concept to illustrate in a plane figure, but is represented by the shaded plane in Fig. 1.03. The transverse radius is usually referred to as ν.

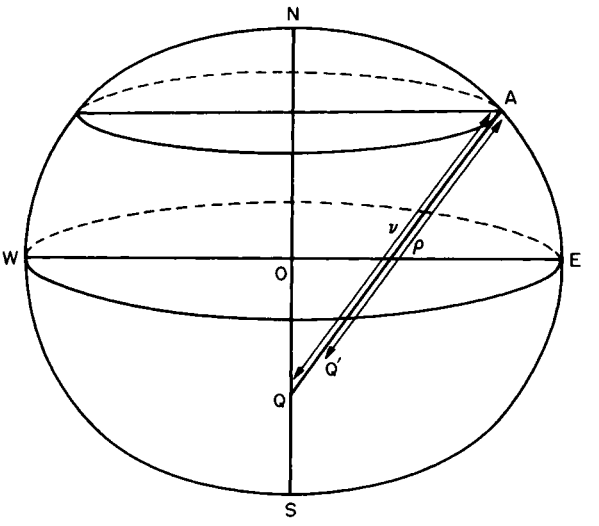


FIG. 4.02 The definition of the meridional radius of curvature (ρ) and the transverse radius of curvature (ν) for a point A on the surface of a spheroid.

From these definitions it follows that both radii lie in the same straight line. The transverse radius is represented by the line AQ, and ends at the point Q on the minor semi-axis in the opposite hemisphere to the point A. The meridional radius is somewhat shorter, as depicted by AQ'. Derivation of the formulae for the two radii of curvature is not attempted here. The two formulae which may be obtained are:

$$\rho = [a(1-e^2)]/[1-e^2 . \sin^2 \varphi]^{3/2} \quad (4.08)$$

$$\nu = a/[1-e^2 . \sin^2 \varphi]^{1/2} \quad (4.09)$$

The two radii have the following properties:

- whatever the latitude $\nu \geqslant \rho$;
- at the poles $\rho = \nu$ and both have their maximum value;
- on the equator, $\nu = a$ and both ρ and ν have their minimum values;
- in latitude 55° or thereabouts, $\rho = a$;
- in latitude 35° or thereabouts, $\rho = b$.

Equations (4.08) and (4.09) indicate that both radii may be completely described in terms of a, e and φ. Since a and e are constants for any particular Figure of the Earth, the only variable is latitude. Most national surveys used to produce geodetic tables containing ρ and ν, together with several other quantities derived from them, for the figure used in that country and for the range of latitude where their activities were concentrated. A few of the older textbooks and manuals also contain shortened versions of the tables. Today, of course, such tables are virtually obsolete. It is so easy to calculate values for ρ and ν by pocket calculator, and this takes less time than was needed to look up and interpolate within the tabulated entries. Since the determination of the radii is usually only a minor stage in more complicated calculations, it is normally done by subroutines to other microcomputer programs.

Arc distances on the spheroid

We deal with the two simple cases first. These are the length of an arc of the equator and the arc of any parallel. On the ellipsoid of rotation both are circular arcs and therefore the simpler geometry of the sphere still applies. The only difference is that the appropriate radius of curvature for the spheroid is used.

On the equator $\varphi = 0°$. Moreover the curvature at right angles to the meridian is the curvature of the equator itself. From the properties listed above,

$$s_e = \delta\lambda . a \quad (4.10)$$

and for the parallel of latitude, φ,

$$s_p = \delta\lambda \,.\, \nu \,.\, \cos\varphi \tag{4.11}$$

Meridional arc distance

The arc of the meridian is more complicated to evaluate because the meridional radius of curvature varies continuously with latitude. Therefore it is necessary to determine, first, the length of a very short arc at a point, and then add together the lengths of all these small elements at all the points which make up the arc.

Let us assume that it is required to determine the arc s_m measured from the equator to latitude φ_1. At any point along this arc we may consider an infinitely small part of it, corresponding to an infinitely small change in latitude $d\varphi$. Within such a small arc distance it is reasonable to state that the arc itself can be regarded as being part of the circumference of a circle. This is shown by Fig. 4.03. Thus for the infinitely short arc we may write (3.12) in the form

$$ds_m = \rho \,.\, d\varphi \tag{4.12}$$

In order to define the length of the whole curve from the equator to a point in latitude φ, it is now necessary to *integrate* the multitude of short arcs which form the whole arc. Since the limits of the arc have already been specified, the arc distance on the ellipsoid, m, may be written as the integral

$$m = \int_{\varphi=0}^{\varphi=\varphi_1} ds_m \tag{4.13}$$

or, from (4.12)

$$m = \int_{\varphi=0}^{\varphi=\varphi_1} \rho \,.\, d\varphi \tag{4.14}$$

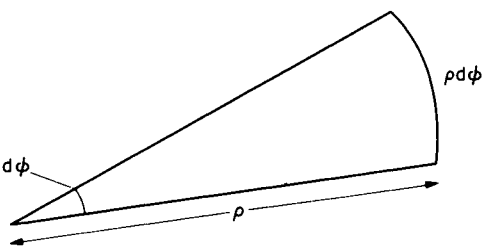


FIG. 4.03 An infinitely short meridional arc element on the spheroid.

and substituting the value for ρ from equation (4.08)

$$m = \int_{\varphi=0}^{\varphi=\varphi_1} [a(1-e^2)]/[1-e^2 . \sin^2 \varphi]^{3/2} . d\varphi \qquad (4.15)$$

after integration of this expression, one version of the equation can be simplified to a form suitable for calculation

$$m = a(1-e^2) . \{A . \varphi - \tfrac{1}{2}(B . \sin 2\varphi) + \tfrac{1}{4}(C . \sin 4\varphi) - \ldots\} \qquad (4.16)$$

where φ_1 is expressed in radians.

In this equation the coefficients A, B, C and D are expressed in terms of e^2 as

$$A = 1 + 3e^2/4 + 45e^4/64 + 175e^6/256 + \ldots \approx 1.0051092 \qquad (4.17)$$

$$B = \quad 3e^2/4 + 15e^4/16 + 525e^6/512 + \ldots \approx 0.0051202 \qquad (4.18)$$

$$C = \quad 15e^4/64 + 105e^6/256 + \ldots \approx 0.0000108 \qquad (4.19)$$

The numerical values for A, B and C have been calculated for the Clarke (1866) spheroid.

We do not expect the reader to undertake the algebraic steps which occur between equations (4.15) and (4.16) without further assistance. This may be found, for example, in Clark (Clendinning) (1944) and many other textbooks on geodesy and surveying. Notwithstanding this obvious short-cut, it is useful to indicate here the initial steps in the argument together with the end-result, and omit most of the intervening stages. Even this abbreviated account demonstrates the greater difficulty encountered in solving a problem on the ellipsoid compared with the trivial calculation for the meridional arc on the sphere in (1.02). Moreover the spheroidal solution is not exact. Equation (4.16) has been terminated at the term (1/4) C . sin 4φ, but could have been extended to include additional coefficients such as D and E. Each of the expressions for A through C are terminated in e^6 but they could have been extended to include the terms in e^8 and e^{10}. However we can see that the numerical value for C in (4.19) is already very small, and that calculation of the additional coefficients and terms in the series would have little effect upon the final result.

The method described above involves expansion of the first eccentricity, e^2 and is therefore sometimes called the *e-series*, for example by Agajelu (1987). An alternative arrangement for terms in e^2 may be found in Clark (1973) and in Snyder (1987a). Similar results can be obtained by forming a series from the parameter e'^2 or n. For example, Williams (1982) makes use of the following expressions:

$$m = \tfrac{1}{2}(a+b)(1+n^2/4+n^4/64+\ldots)(\varphi - \alpha . \sin 2\varphi + \beta . \sin 4\varphi - \gamma . \sin 6\varphi + \delta . \sin 8\varphi - \ldots) \qquad (4.20)$$

where

$$\alpha = 3n/2 - 9n^3/16 + 3n^5/32 - \ldots \quad (4.21)$$

$$\beta = 15n^2/16 - 15n^4/32 + \ldots \quad (4.22)$$

$$\gamma = 35n^3/48 - 105n^5/256 + \ldots \quad (4.23)$$

$$\delta = 315n^4/512 - \ldots \quad (4.24)$$

Agajelu (1987) also offers this *n-series* in a slightly different form.

Meridional arc distance measured between two points of known latitude

Thus far we have only examined the formulae needed to determine meridional arc length from the equator to a point in latitude φ. In many practical calculations the meridional arc length required extends from latitude φ_1 to latitude ϕ_2. Of course this could be obtained by determining m_1 and m_2 separately and subtracting one from the other. However, this is clumsy compared with determining the equation for the developed arc of the meridian between two latitudes. Various formulae have been proposed, the most elegant being that using series in the parameter n and used, for example, in Ordnance Survey (1950). If $\varphi_2 > \varphi_1$ and $\delta\varphi = \varphi_2 - \varphi_1$

$$m_1 - m_2 = b\{(1 + n + 5n^2/4 + 5n^3/4) . \delta\varphi - (3n + 3n^2 + 21n^3/8) . \sin\delta\varphi . \cos\delta\varphi + (15n^2/8 + 15n^3/8) . \sin 2\delta\varphi . \cos 2(\varphi_2 + \varphi_1) - 35n^3/24 . \sin 3\delta\varphi . \cos 3(\varphi_2 + \varphi_1)\} \quad (4.25)$$

In many applications the lower latitude φ_1 is that of the origin of the projection in use. For example, solving this equation for the length of the developed arc of the meridian used in the Ordnance Survey Transverse Mercator projection, we put $\varphi_1 = 49°$, this being the latitude of the origin of this projection as well as being, as we have already seen, the true origin of the National Grid.

Numerical solutions from series expansions

Some explanation about how such expressions are derived is necessary.

They are based first upon the replacement of the terms in equation (4.15) by their integrals, and secondly, for greater ease of both analysis and computation, by the expansion of these in series. This procedure is well known in elementary calculus, where Taylor's and Maclaurin's Theorems may be used to obtain a series corresponding to any specified function. For example, we may convert the function sin x into a series of

terms containing ascending powers of the variable x, expressed in radians. Thus

$$\sin x = x - x^3/6 + x^5/120 - x^7/5040 + \ldots \quad (4.26)$$

This equation is useful from both the practical and algebraic points of view. It is the method which is used to obtain the numerical values of trigonometric functions published in tables and used by the subroutines of digital computers to evaluate all the standard functions which are available in a particular instrument. Inspection of equation (4.26) indicates that the right-hand side of it is composed of four terms in ascending powers of x. The numerical value in the dominator of each fraction represents the factorial (!) of a number. For example, in the second term, $6 = 3! = 1 \times 2 \times 3$.

Since the numerical value of x lies within the range $x = 0$ to $x = \pi/2 = 1.57\ldots$, the values of x^3, x^5 etc. increase more slowly than their respective denominators. Consequently *the size of the terms on the right-hand side of the equation decreases from left to right.* Since $x^7/5040 < x^5/120 < x^3/6$ the series may be said to *converge*. The right-hand side of equation (4.20) could be further extended to include terms in x^9 and so on, but the effect of these terms upon the numerical values of sin x would be negligible up to the sixth decimal place.

Snyder (1987a) has listed three types of such expressions which arise commonly in mapping from the spheroid. We shall encounter these later.

The calculation of the length of any arc and its azimuth on the spheroid

Since the simple meridional arc introduces such difficulties it is not surprising that the determination of the length and azimuth of any arc is even more complicated. Yet this kind of problem arises in control surveys, either in the sort of work which extends through a big country or for extremely precise work in a smaller area. In other words it arises in those cases where the spheroidal assumption is mandatory. One problem is to find the geographical coordinates of a new survey station from those of a point already fixed, using the measured or calculated bearing and distance to the new station. The converse problem, which is less common, is to determine the bearing and distance between two stations from their known geographical coordinates. In mathematical geodesy a variety of different formulae have been described. These are usually referred to by the name of their originator, such as *Clarke's Formula for Long Lines*, *Puissant's Formula*, *Rainsford's Extension of Clarke's Approximate Formula*. Bomford (1962) discusses the merits and accuracies of ten such formulae.

In the days before digital processing was easily accessible, these geodetic

computations were a headache to the field surveyor who might need to do the calculations with no more than an adding machine and a volume of logarithmic functions to assist him. For this reason many short-cut methods were developed to simplify the problem of calculation. The most useful modifications were those originally adopted by Gauss for his *Mid-latitude Formula*, and by Clarke in his celebrated formula for short and medium-length lines. Both make use of the idea that if an *auxiliary sphere* be fitted tangentially to the surface of the ellipsoid, near the middle of an arc, or in the middle latitude between two points, these surfaces do not depart appreciably from one another over distances of a few tens of kilometres. Hence the computations make use of the radii of curvature of the spheroid, but the methods of spherical trigonometry to solve the triangles.

Arc distance defined in three-dimensional cartesian coordinates

Nowadays the solution for the length of an arc is most likely to be obtained through the determination of the differences between the three-dimensional cartesian coordinates of the terminal points using the following argument and equations. The method is based upon the determination of the straight line chord distance between the terminal points, to which a chord-to-arc correction is applied to find the length of the curve.

The first stage in the determination is to find the *reduced* or *parametric latitude*, u, of each point from the equation

$$\tan u = (b/a) \tan \varphi \tag{4.27}$$

This angle differs from geocentric latitude already defined because it measures the angle to a point A' lying on the surface of the auxiliary sphere illustrated in Fig. 4.04. This sphere is tangential to the equator and therefore has radius $R = a$.

The point A on the spheroidal surface may be expressed in three-dimensional cartesian coordinates as

$$\begin{aligned} X_A &= a \,.\, \cos u \,.\, \cos \delta\lambda \\ Y_A &= c \,.\, \cos u \,.\, \sin \delta\lambda \\ Z_A &= b \,.\, \sin u \end{aligned} \tag{4.28}$$

If we express the corresponding coordinates of a point B as X_B, Y_B and Z_B respectively, the coordinate differences between A and B may be

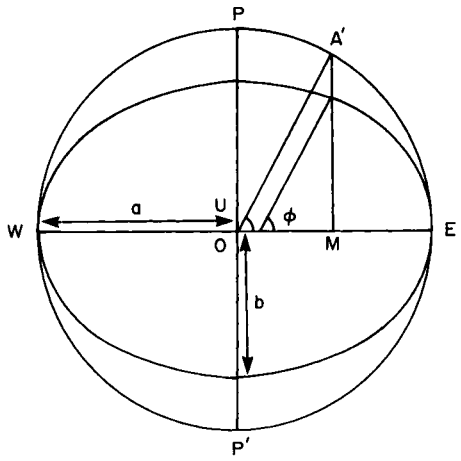


FIG. 4.04 The parameters used to determine the three-dimensional cartesian coordinates of a point on the surface of a spheroid as a preliminary to the determination of the length of the arc and the bearing between two points. Definition of the *reduced* or *parametric latitude*, u.

written

$$\begin{aligned} \delta X &= X_A - X_B \\ \delta Y &= Y_A - Y_B \\ \delta Z &= Z_A - Z_B \end{aligned} \tag{4.29}$$

The exact chord distance, K, between A and B may now be determined from

$$K^2 = \delta X^2 + \delta Y^2 + \delta Z^2 \tag{4.30}$$

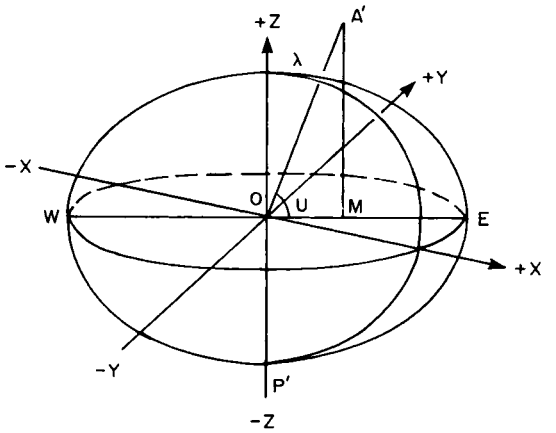


FIG. 4.05 The three-dimensional coordinate system used to calculate the position of the point $A = (X,Y,Z)$ on the curved surface of the spheroid.

It is now necessary to apply a correction D–K in order to convert from chord into arc distance for this varies with the length and the direction of the line. This is not an exact solution, but introduces a small amount of approximation. The required expression is

$$(\text{D–K}) = \text{K}^3/24\text{R}^2 + 3\text{K}^5/640\text{R}^4 \qquad (4.31)$$

where R is the mean radius of the spheroidal arc. This has to be computed from *Euler's Theorem*

$$\text{R} = \rho\nu/(\rho \sin^2\alpha + \nu\cos^2\alpha) \qquad (4.32)$$

where α is the bearing, or azimuth of the line from A to B. This, in turn, has to be found from the equation

$$\tan\alpha = \sin\delta\lambda/\{\sin\varphi_a[\cos\delta\lambda - (\cos u_a/\cos u_b)(1-b^2/a^2)] - b^2/a^2\tan\varphi_b\cos\varphi_a\} \qquad (4.33)$$

For lines which are less than 100 km in length the correction (D–K) is less than 1 m (or $< 0{\cdot}001\%$); at 350 km, the correction is about 44 m ($\approx 0{\cdot}013\%$). For a line of length 1000 km the correction is of the order of 1030 m, which is about 0·1% of the length of that line.

The inverse transformation from spatial into geographical coordinates has been described by Bowring (1976).

The choice of radius for an auxiliary sphere

Equation (4.32) introduces once more the consideration of a suitable value for R which, in this case, is the radius of the auxiliary sphere used to determine the arc length. Consequently we must return to the discussion of the types of reference figure which may be used and the purposes for which they are needed. We have already emphasized, on p. 61, the need for a critical approach in the choice of suitable metric dimensions for R. We have argued that since the spherical assumption involves some approximation, the use in the general study and construction of map projections of a value for R which has been calculated to the nearest metre, represents a needless degree of accuracy.

There is also a variety of maping activities where the opposite prevails and unjustified degrees of approximation may be involved. An example of this is the use of equation (3.20) to determine the lengths of a great circle arc combined with a supposedly precise value for R in order to convert from angular into linear measure. This practice has arisen in that rather grey area of activities which belongs to navigation, hydrographic surveying and geodetic surveying and which characterised some of the early offshore activities of seabed exploration together with the legal work of defining maritime boundaries. For example, the distance measurements

used to locate median line between Britain and Norway in the northern North Sea were originally based upon the spherical assumption but now have to be used as if they were geodetically precise. The present author, in Maling (1989) has already been at pains to demonstrate the unsuitability of the spherical assumption for this kind of work, and has argued for the correct application of rigorous computations on the spheroid. We shall find, moreover, that these are only two aspects of a larger problem and for other purposes we must still, nevertheless, define R more accurately.

Between the spheroid and the plane representation of it as a map projection we may have recourse to an auxiliary sphere, either to facilitate geodetic computations, as in the example described above, or in the use of this intermediate body for methods of *double-projection*, this being two-stage mapping of the spheroid, first from spheroid to sphere and secondly from sphere to plane. This subject is examined in greater detail in Chapter 16.

It is a property of the auxiliary sphere that it is considered to be tangential to the spheroid at some suitable place. Many of the following definitions depend upon this choice.

(1) The simplest choice for the radius of an auxiliary sphere is to use *one of the semi-axes of the ellipsoid*, or some combination of both of them to provide a single value. All of the following have been used

- The equatorial radius of the spheroid. This corresponds to the use of the major semi-axis, a, so that the sphere is tangential to the spheroid at the equator, as illustrated by Fig. 4.04.
- The combination of both semi-axes a and b. This may be either the arithmetic mean:

$$R = \tfrac{1}{2}(a+b) \tag{4.34}$$

or the geometric mean:

$$R = \sqrt{(ab)} \tag{4.35}$$

- In a triaxial ellipsoid the equator must also be defined by two axes. If, therefore, the Figure of the Earth were to be regarded as a triaxial ellipsoid it would be necessary to take three axes into consideration, whereas in an ellipsoid of rotation the equator is a circle so that, as we have seen, the body is completely defined by only two axes. In order to make approximation of the triaxial body it is therefore usual to take twice the value of the major semi-axis to give a radius which is either

$$R = (2a+b)/3 \tag{4.36}$$

which is the arithmetic mean, or

$$R = (2ab)^{1/3} \tag{4.37}$$

which is the geometric mean.

(2) We may use *the radii of curvature of the spheroid for some reference latitude.* For example,

- The value of ρ or v may be taken for the latitude 45°, this being chosen because this latitude lies midway between the equator and the poles and corresponds to an auxiliary sphere which intersects the spheroid in this latitude.
- More commonly, the values for ρ or v are taken for the mean latitude, φ_m, of an arc, a zone or a quadrangle formed by two meridians and two parallels. Then we calculate the radii of curvature for this latitude and use ρ_M or v_M.
- Again we may use either the arithmetic or geometrical means of the radii of curvature of part of the surface, using

$$R = (\rho + v)/2 \tag{4.38}$$

or

$$R = \sqrt{(\rho v)} \tag{4.39}$$

respectively. The last of these is known as the 'Gaussian Curvature'.

(3) The radius may be determined for *a sphere having the same volume as the chosen figure of the earth.* This radius may be determined from the expression

$$R = a(1 - f/3 - f^2/9) \tag{4.40}$$

where f is the flattening of the spheroid.

(4) The radius may be determined for *a sphere having the same surface area* as that of the chosen Figure of the Earth. This is also known as the *Authalic Sphere* and

$$R^2 = (a^2/2\pi) \cdot \{1 + (1 - e^2/2e) \cdot \ln[(1+e)/(1-e)]\} \tag{4.41}$$

(5) The *rectifying sphere* has meridional length equal to that of the spheroid. Adams (1921) derived it as

$$R = a(1-n)(1-n^2)(1 + 9n^2/4 + 225n^4/64 + \ldots) \tag{4.42}$$

(6) The solution derived from Euler's Theorem, which refers to an arc of specific length and azimuth, has already been listed in equations (4.32) and (4.33).

The variability of R resulting from so many different definitions gives rise to markedly different values for the spherical distance. The methods

TABLE 4.02 *The values of spherical radius, R, determined from the International Spheroid by different methods*

Definition of spherical radius	R (metres)
Major semi-axis, a	6 378 388
Arithmetic mean of a and b	6 367 650
Geometric mean of a and b	6 367 641
Arithmetic mean of three axes $(2a+b)/3$	6 371 229
Geometric mean of the three axes $(2a \cdot b)^{1/3}$	6 371 221
g . m for $\varphi = 15°$	6 359 778
ρ for $\varphi = 45°$	6 367 586
ν for $\varphi = 45°$	6 389 135
g . m $\varphi = 45°$	6 378 351
ρ for $\varphi_m = 60°$	6 383 727
ν for $\varphi_m = 60°$	6 394 529
g . m $\varphi_m = 60°$	6 389 126
Sphere of equal volume	6 371 221
Authalic sphere	6 371 228
Rectifying sphere	6 367 655
Range	34 751

outlined above have been used to determine values of R which are based upon the International Spheroid (1924). The results are listed in Table 4.02. The range in this table is nearly 35 km, which is too big to be ignored even when using R to construct small-scale atlas maps.

CHAPTER 5

Some basic ideas about the mathematics of map projections

> This is why Elastoplast which stretches is a better fit than ordinary Elastoplast for cuts on knuckles and knees.
>
> Jeremy Gray, *Ideas of Space*, 1979

Introduction

A map projection may be defined as: * *any systematic arrangement of meridians and parallels portraying the curved surface of the sphere or spheroid upon a plane*. For many purposes in the present book, *it will suffice to regard the earth as a perfect sphere*. This has the advantage of being mathematically simpler to understand without losing sight of any of the salient problems which have to be tackled. The main exceptions to the use of the spherical assumption come in Chapters 15, 16 and 19, where the specialised uses of projections in surveying and topographical cartography are considered.

It was stated in Chapter 2 (p. 28) that every map projection is a form of coordinate representation upon the plane, and that its graticule intersections may be located by means of either cartesian or polar coordinates. In other words, each point on the earth's surface, with geographical coordinates (φ, λ) may be reproduced on the plane by a point located in either the (x, y) or (r, θ) systems of plane coodinates.

Functional relationships

We may express this idea in the generalized form of *functional relationships* (or *functions*) and write

$$x = f_1(\varphi, \lambda) \tag{5.01}$$

$$y = f_2(\varphi, \lambda) \tag{5.02}$$

or

$$r = f_3(\varphi, \lambda) \tag{5.03}$$

$$\theta = f_4(\varphi, \lambda) \tag{5.04}$$

These expressions are the mathematical shorthand for statements such as 'x is a function of latitude and longitude', etc. The suffices f_1–f_4 indicate that these are different functions. Thus we may distinguish between (5.01) and (5.02) by the statement 'whereas x is one function of both latitude and longitude, y is a different function of these variables'. We can further state that in (5.01), x is the *dependent variable* which is a function of two *independent variables*, φ and λ.

At this stage we do not precisely specify the nature of these functions. Each map projection has unique equations for x and y or r and θ, which will be used to define and construct it. Appendix I, on pp. 430–441, indicates some of the formulae which these functions represent. For the present, however, the generalized expressions of (5.01)–(5.04) are useful for the preliminary study of the subject, for they indicate certain important relationships between the sphere and the plane. Moreover, they serve as a convenient basis for the systematic classification of all map projections.

From the statement that x and y (or r and θ) are functions of latitude and longitude it follows that one point (φ, λ) on the earth is represented by one point (x, y) or (r, θ) on the map. In other words there is a *one-to-one correspondence* between the earth and the map. We will have to qualify this statement later because some map projections show the same meridian twice, because the geographical poles are represented by lines instead of by points, or because certain parts of the earth's surface cannot be shown on the projection. These peculiarities arise from the simple fact that a sphere has a *continuous surface* whereas a plane map must have a boundary. The kinds of peculiarities which have been mentioned generally occur at the edge of a map projection and they must be considered to be exceptional, or *singular points*. Within the body of the majority of map projections each point on the earth is shown only once; therefore the idea of corresponding points holds good.

The correspondence between points on the surface of the earth and the plane map cannot be exact. In the first place, some kind of scale change must occur because a map of the earth at scale 1/1 is a physical impossibility. Secondly, the curved surface of the earth cannot be fitted to a plane without introducing some *deformation* or *distortion* which is equivalent to stretching or tearing the curved surface.

Principal scale

Because a map is a small-scale representation of the earth it is necessary to consider this part of the transformation first.

In the everyday meaning of the word, *scale* may be defined as: * *the ratio of distance on a map, globe or vertical section to the actual distances they represent.* Expressed geometrically, if the map distance is $A'B'$, corresponding to the ground distance AB, the scale of the map is the fraction $A'B'/AB$, expressed as a fraction whose numerator is 1. Thus, if 40 mm on the map corresponds to 1 km on the ground, $A'B' = 40$ and $AB = 1000 \times 1000 = 1\,000\,000$ (to bring AB into the same units as $A'B'$) and the scale 40/1 000 000 may be described by the *representative fraction* 1/25 000.

Generating globe

From the definition of scale given above, precisely the same reasoning may be used to describe the scale of a globe used to represent the earth. In this case, comparison is made between the lengths of two corresponding arcs of great circles, AB on the earth and $A'B'$ on the globe. From equation (1.02) and the arguments presented in Chapter 3:

$$AB = \mathrm{R} \,.\, \mathrm{z}$$

and

$$A'B' = \mathrm{r} \,.\, \mathrm{z}$$

Hence the scale of the globe may be expressed as

$$A'B'/AB = (\mathrm{r} \,.\, \mathrm{z})/(\mathrm{R} \,.\, \mathrm{z}) \tag{5.05}$$

or

$$1/S = \mathrm{r}/\mathrm{R} \tag{5.06}$$

where S is the denominator of the representative fraction, r is the radius of the globe and R is the radius of the earth. For example, a globe of radius 212 mm will have a scale denominator

$$S = 6\,371\,100/0{\cdot}212$$

$$= 30\,052\,358 \approx 30\,000\,000$$

so that the globe evidently has scale 1/30 000 000.

We assume that *generating globe* is an exact replica of the earth but at the scale indicated by (5.06). We call this the *principal scale* and therefore can define it as: * *the scale of a reduced or generating globe representing the sphere or spheroid defined by the fractional relation of their respective radii.*

The concept of a generating globe of known principal scale is extremely useful in the discussion which follows. Since a map is a small-scale representation of the whole or part of the surface of the earth we are

accustomed to think of all matters relating to scale in terms of representative fractions like 1/30 000 000. Since we must consider in detail how the transformations from sphere to plane can be accomplished and, in particular, investigate how and where distortion in scale may occur, it is inconvenient to have to think always in these terms and regard the scale changes as, for example, between 1/30 000 000 and 1/29 500 000, or even worse, the differences between $3{\cdot}333^{-8}$ and $3{\cdot}389^{-8}$. We therefore sweep away this difficulty by thinking in terms of the generating globe which is a replica of the earth at the scale of the map. Since we wish to eliminate the use of awkward fractions altogether, we define the principal scale as

$$\mu_0 = 1{\cdot}0 \tag{5.07}$$

and evaluate distortion as some multiple of this.

It follows, moreover that *the principal scale is equivalent to the representative fraction printed in the margin of the map.* Hence we have the statement that

$$1/S = \mu_0 = 1{\cdot}0 \tag{5.08}$$

Introduction to the concepts of distortion

At the manageable dimensions of a generating globe it is easy to demonstrate that the curved surface of a sphere cannot be fitted to a plane. This fundamental principle can be verified easily by anyone who experiments with a globe, beach ball or similar smooth surface. If we attempt to fit a small piece of paper – a postage stamp, for example – to the surface of a large beach ball, it is possible to make it adhere without creating any wrinkles or tears in the paper. This is because the piece of paper is small compared with the ball and the deformation of the plane which is needed to make the two surfaces fit is less than can be accommodated by the elasticity of the paper. On the other hand, the same postage stamp cannot be fitted to the curved surface of a table tennis ball without the introduction of considerable folding, tearing or creasing.

An important conclusion to be derived from these simple experiments is that if the area of the plane surface is small compared with the total surface area of the sphere, the amount of distortion introduced is less than occurs when the area of the plane corresponds to a larger part of the curved surface. This is a qualitative, empirical observation similar to that made in Chapter 1, p. 20, with reference to the use of the assumption that the earth is a plane surface. However, it is now important for us to learn more about these processes of distortion and, in particular, discover how they may be expressed algebraically and used quantitatively to illus-

trate how a particular map projection distorts the curved surface of the globe.

Of course the experiments with a ball and postage stamp are the converse of the object of creating a map projection, which is to make parts of the curved fit a plane. A useful illustration, which may be simulated by cutting orange peel and laying this flat, is to imagine that the curved surface of a globe has been cut along certain parallels and meridians, as shown in Fig. 5.02. If the spherical surface is cut thus it is very nearly possible to lay it flat. However, this result is obtained only at the expense of showing certain parallels of latitude twice, and interrupting the continuity of the map by leaving gaps between these parallels. If it is desirable to map the whole surface continuously, these gaps must be closed by stretching each zone in a meridional direction until the corresponding parallels meet, as illustrated in Fig. 5.02. Stretching of the map involves distortion, and comparison of Figs 5.01 and 5.02 indicates that the amount of stretching increases progressively towards the edges of the map. The amount of distortion may be indicated by the deformation of the circles shown in Fig. 5.01 into the oval figures shown in Fig. 5.02.

In the creation of the continuous map illustrated by Fig. 5.02 the distortion described is *linear distortion* directed along the meridians. The graphical result is that the distance between any two parallels of latitude increases from the middle of the map towards its edges. On the other hand, the distances between successive meridians vary only with latitude. We have already seen that this is a property of the spherical surface. Equation (3.17) describes it. If, however we consider the spacing between the meridians along any particular parallel of latitude, we see that it is almost constant and equivalent to the spacing illustrated in Fig. 5.01. This suggests that linear distortion in this projection occurs in one direction but not in the other. This is clearly likely to influence the representation of both *angles* and *areas* on the map. The effect may be demonstrated by drawing two simple diagrams, as shown in Figs 5.03 and 5.04. In Fig. 5.03 the point P has coordinates (10,10), measured in a system with origin O. It follows that the angle $YOP = 45^\circ$ and the area of the square $YOXP = 100$ square units. In Fig. 5.04 the scale along the ordinate has been doubled but that along the abscissa remains unchanged. Thus $P' = (10,20)$. The angle $Y'OP' = 30^\circ$ and the area of the rectangle $Y'OX'P' = 200$ square units. We will call the change in angle $Y'OP - YOP$ the *angular deformation* and the change in area $Y'OX'P' - YOXP$ the *exaggeration of area*. In a map projection they are not as easily defined as they are in a pair of plane graphs, but the essential characteristic is clear. *Both angular deformation and exaggeration of area depend upon linear distortion and therefore they may be defined in terms of this.* Consequently it is the change which occurs in the length of any line which is fundamentally important to the study of map projections.

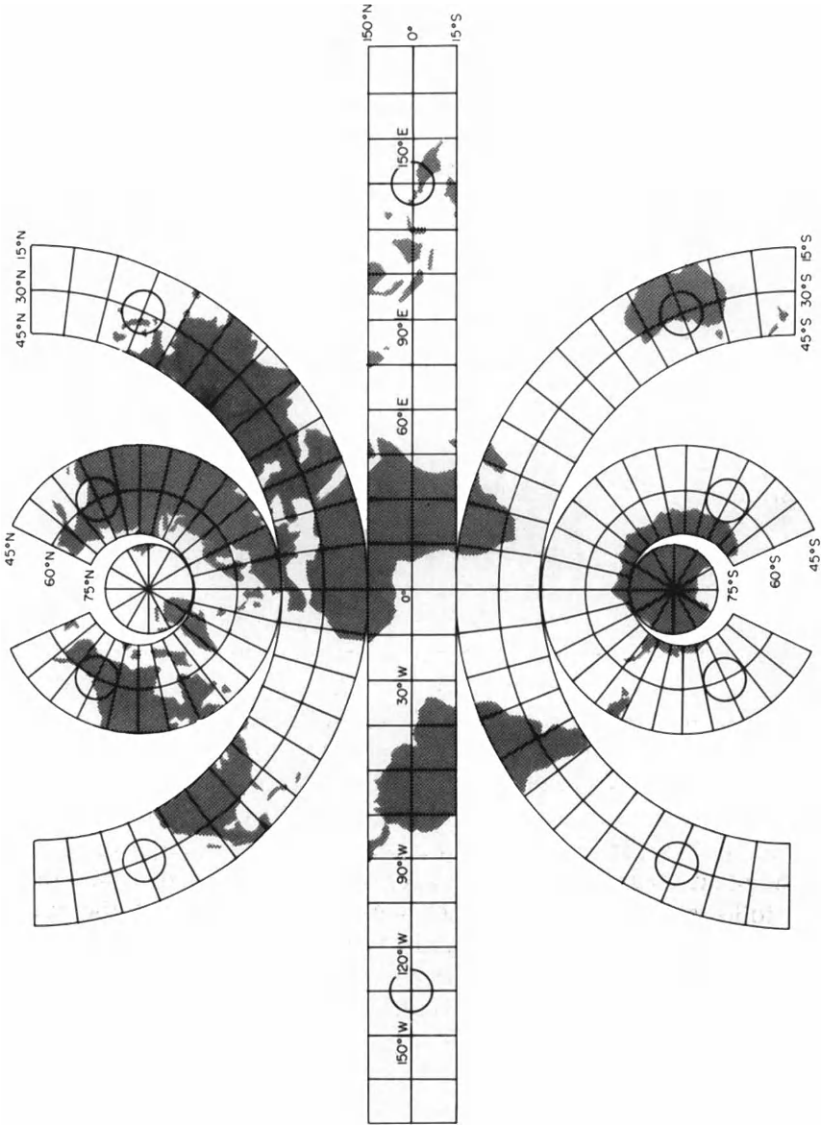


FIG. 5.01 Plane representation of the curved surface of the earth obtained by cutting the spherical surface along the parallels 15°N and S, 45°N and S, 75°N and S and along the antimeridian of Greenwich.

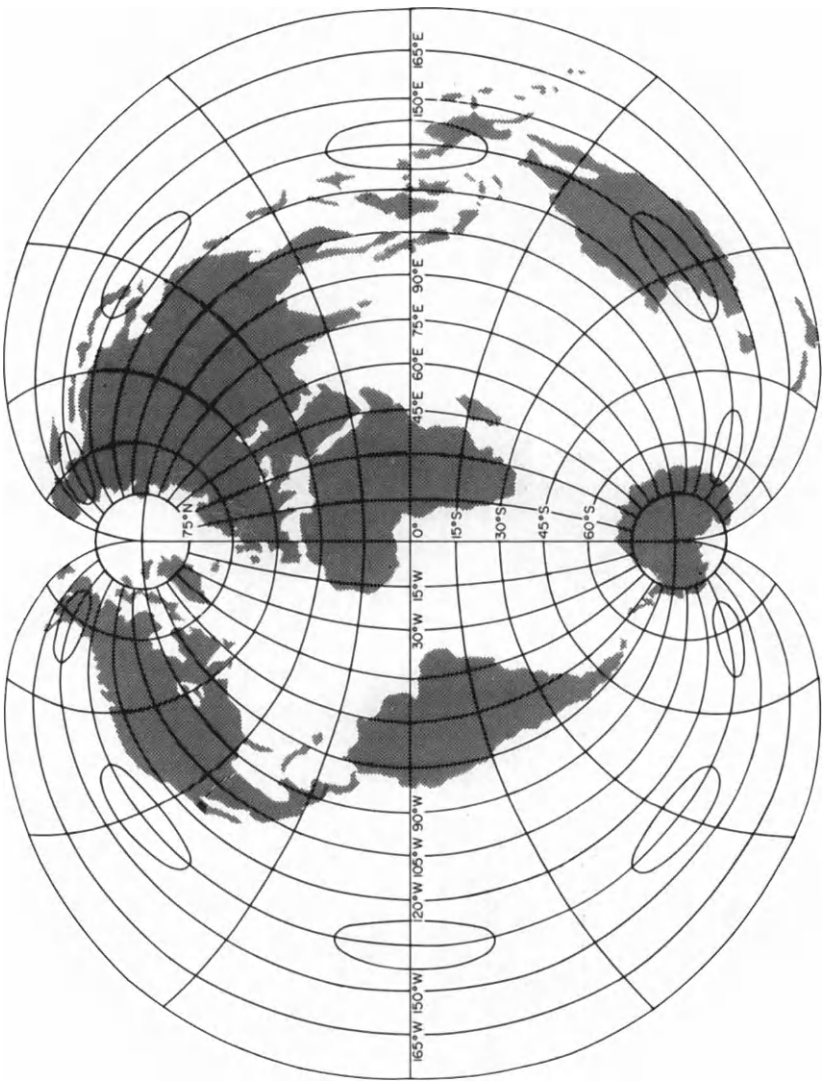


FIG. 5.02 Continuous plane representation of the earth derived from Fig. 5.01 by meridional stretching. Note how the circles shown in Fig. 5.01 have been deformed into ovals on this diagram. The map is the world development of the Polyconic projection (No. 38 in Appendix I) in which the parallels have equidistant separation along each meridian.

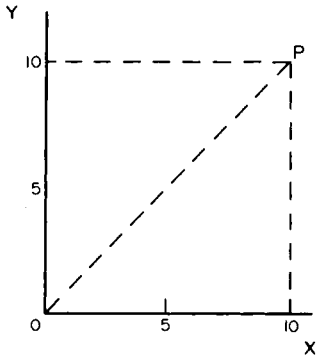


FIG. 5.03 Demonstration of the influence of linear distortion upon angular and area representation. Stage 1, initial condition where $P = (10,10)$.

Linear distortion

When the scale of a map is known from its representative fraction, one might suppose that this scale is constant in three respects:

(1) That *the ratio established by the representative fraction applies to the lengths of all lines measured on the map.* For example, if the scale of

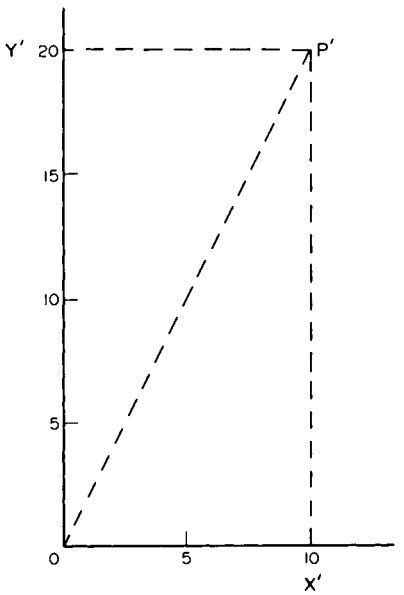


FIG. 5.04 Demonstration of the influence of linear distortion upon angular and area representation. Stage 2, showing the result of changing the ordinate so that $P' = (10,20)$.

the map is 1/25 000, we expect that a line of length 40 mm corresponds to a ground distance of 1 km. We may further assume that a line of length 80 mm corresponds to a line of length 2 km and that a line of length 400 mm corresponds to 10 km on the ground. Hence we may assume that the relationship established by the representative fraction is constant for linear measurement of any distance which can be contained within the neat lines of the map. Moreover we assume that the same relationship will hold good for all maps of the same scale irrespective of the part of the world which they depict.

(2) That *the relationship established by the representative fraction is constant for all parts of the map.* Thus we suppose that a line of length 40 mm corresponds to a ground distance of 1 km, whether this be measured in the centre of the map or near one edge or corner of the sheet.

(3) That *the relationship is also independent of direction.* Thus, at 1/25 000 scale, 40 mm represents 1 km irrespective of whether the line to be measured lies north–south or east–west or in any intermediate direction.

These three assumptions appear to be axiomatic in most kinds of map use, to the extent that the majority of map or chart users apply them without further thought. However, *the assumption that scale is constant for all distances, at all places and in all directions is not true.* If it were possible to reproduce the principal scale in all directions and everywhere upon the plane surface of the map, then the map would be a perfect representation of the spherical surface and therefore it would be part of the spherical surface. Since a curved surface is not a plane it follows that the transformations to the plane *must* involve variation in scale in some or all of the three ways which have been specified.

The numerical example refers to a map of scale 1/25 000 which probably represents a good area of 100–200 square kilometres. Within this small portion of the earth's surface the scale changes are small; so small that negligible errors are introduced by making the assumption that scale is constant. The errors are much less than the uncertainty in position caused by representing ground detail by legible lines of exaggerated width; they are also less than the variations in paper size and shape which occurs with changes in humidity and temperature. But it is important to realize that linear distortion is still there, even if it is too small to be measured or recognised by our rather crude methods.

Lines and points of zero distortion

Although it is clearly impossible to create a perfect map in which the principal scale is preserved everywhere, it is quite easy to maintain the principal scale along certain lines or at certain points on the map. Along

these lines, or at these points, scale is constant and equal to the principal scale so that no linear distortion is present. Thus we have the following terms and their definitions:

1. * *Line(s) of zero distortion are lines on a map projection along which the principal scale is preserved and which correspond to certain great circle or small circle arcs on the sphere.*
2. * *A point of zero distortion is a point on a map projection where the principal scale is preserved.*

The meanings of these definitions may be demonstrated by some well-known experiments with a globe or ball and a sheet of paper. These are illustrated in Figs 5.05, 5.06, 5.07, 5.08, 5.09 and 5.10. We use the paper to create a cylinder, cone or plane. The first and second of these are *developable surfaces*, these being surfaces that can be transformed into a plane without distortion.

If the sheet of paper is wrapped round the sphere in the form of a cylinder, it makes contact with the spherical surface along the circumference of a great circle, as illustrated in Fig. 5.05. By marking the paper we see that the length of the line of contact on the plane sheet, unrolled from the cylindrical form, is the same as the length of the circumference of the great circle.

The second possibility is to wrap the sheet of paper in the form of a cone (Fig. 5.06) so that this surface makes contact with the spherical surface along the circumference of a small circle. Again it is obvious that the length of the line of contact between the paper cone and the globe

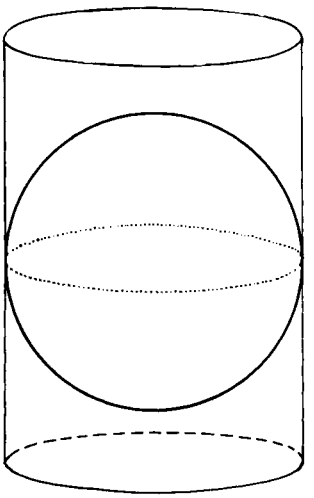


FIG. 5.05 The tangent cylinder.

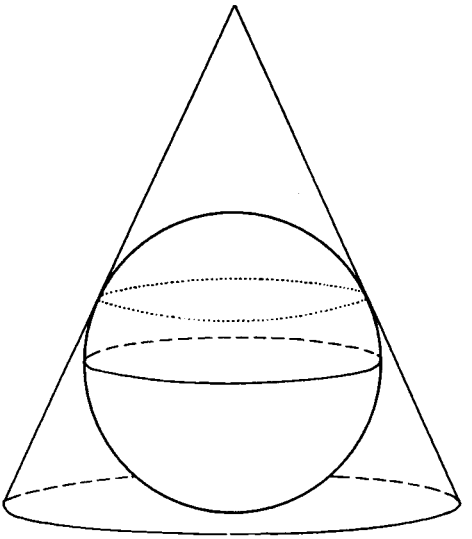


FIG. 5.06 The tangent cone.

corresponds to the length of the circumference of the small circle. The third possibility is to hold the paper as a plane surface so that it forms a tangent plane to the globe (Fig. 5.07). Although it cannot now be demonstrated that lines of finite length are represented at true scale, it follows from the definition of a spherical angle (p. 53) that any angle drawn on the plane at the point of contact is equal to the corresponding spherical angle on the globe.

We may also consider three analogous cases where the surface of the developable surface intersects the surface of the globe. These cannot be simulated by experiment with a sheet of paper but are easy enough to illustrate. Figure 5.08 shows the *secant cylinder* which intersects the sphere

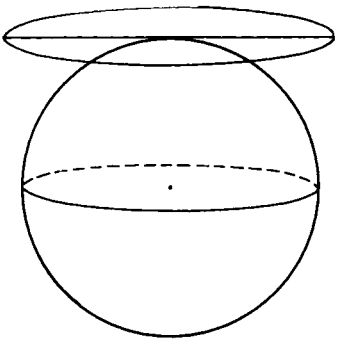


FIG. 5.07 The tangent plane.

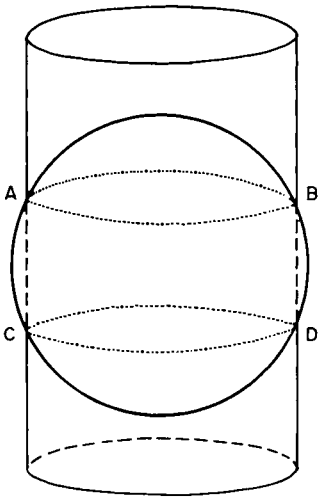


FIG. 5.08 The secant cylinder.

along two arcs of small circles, *AB* and *CD*. It is easy to demonstrate from the geometry of a sphere that since a cylinder has constant radius, the small circles have the same radii and therefore they are equidistant from the plane of the great circle defined by the co-axial tangent cylinder. By reasoning analogous to that followed for the tangent cylinder, the principal scale is preserved along the circumferences of both small circles.

The example of the secant cone is illustrated in Fig. 5.09, where it can be seen that two small circles of different radii are defined by *AB* and *CD*, and each of them is represented on the cone at its correct length. In Fig. 5.10 the tangent plane has been displaced so that it now intersects the spherical surface and the small circle *AB* is traced upon this plane. It follows from the definition of a small circle that the circumference traced on the plane is identical with the circumference of it on the sphere.

The experiments and illustrations which depict the various ways in which the location of lines or a point of zero distortion may be imagined indicate that the lengths of lines should be the same as those on the generating globe. It is less easy to demonstrate that this principle applies also with the infinitely small circle centred on a point of zero distortion. The main difficulty is to imagine how we can define scale at a point. We have to reconcile the elementary concept of scale as a fraction relating finite distances whereas the Euclidean definition of a point is that it has position but no magnitude. To proceed further necessitates some reconsideration of the concept of scale in terms of the differential calculus and determine the rate at which scale may change along a line which is infinitely short.

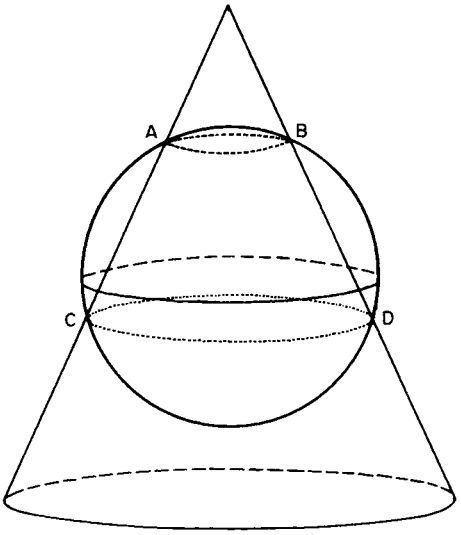


FIG. 5.09 The secant cone.

Particular scales

Let us consider a map projection of part of the surface of a globe which satisfies equations (5.01) and (5.02). In other words we define positions on the plane in rectangular coordinates (x, y) and we know that the positions of points plotted within this system are some function of both latitude and longitude. Thus if φ changes, both x and y are altered. Similarly if λ changes both x and y are altered. Figure 5.11 represents part of the curved surface of the generating globe and shows the spherical quadrilateral (or *quadrangle*) formed by the intersection of a pair of meridians by a pair of parallels. We assume that the geographical coordinates of the point A are (φ, λ) and that the other three points, B, C and

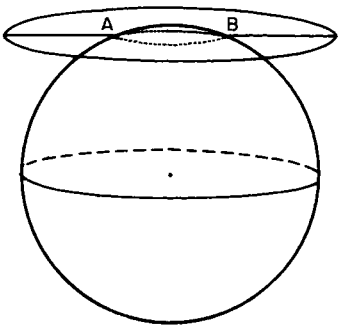


FIG. 5.10 The secant plane.

TABLE 5.01

Point	Latitude	Longitude
A	φ	λ
B	$\varphi+\delta\varphi$	λ
C	$\varphi+\delta\varphi$	$\lambda+\delta\lambda$
D	φ	$\lambda+\delta\lambda$

D lie to the north and east of A. Then, denoting the difference in latitude between the parallels as $\delta\varphi$ and the difference in longitude between the meridians as $\delta\lambda$ we may list the geographical coordinates of the four points according to the system shown in Table 5.01. Figure 5.12 shows a map of the corresponding points A', B', C' and D'. The rectangular coordinates of the point A' are (x, y) and the coordinate differences between A' and C' are δx and δy.

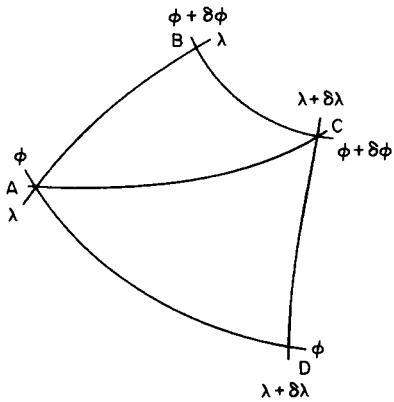


FIG. 5.11 A spherical quadrilateral $ABCD$ of finite size formed by the intersections of the parallels φ and $\varphi+\mathrm{d}\varphi$ with the meridians λ and $\lambda+\delta\lambda$.

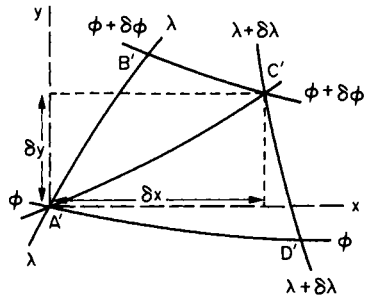


FIG. 5.12 The plane representation $A'B'C'D'$ of the spherical quadrilateral illustrated by Fig. 5.11.

Thus far we have regarded the spherical quadrilateral and its projection as having finite size. In other words the quantities $\delta\varphi$, $\delta\lambda$, δx and δy can be measured on a globe or map. Because we have specified a generalized functional relationship, the sides and diagonal of $A'B'C'D'$ may be composed of curves, and the angles between these sides may be of any size.

Differential geometry of the sphere and plane

In order to proceed further with the analysis it is now necessary to consider that the corresponding figures have been reduced in size until they are infinitely small. This has two important consequences:

- *the shapes of corresponding lines on both globe and map approximate more and more closely to straight lines;*
- *the angles formed by the intersections of pairs of lines remain unchanged.*

It follows that the spherical quadrilateral formed originally by pairs of meridians and parallels intersecting at right angles is transformed into a rectilinear figure in which all four angles are still right angles. Hence in Fig. 5.13, $ABCD$ is a rectangle. On the map the sides and diagonals of the figure $A'B'C'D'$ are transformed into straight lines, but angles such as θ' are preserved. Figure 5.14 illustrates this transformation in enlarged form. We regard the points A and A' as having the coordinates already allocated to them, but denote the incremental changes in latitude, longi-

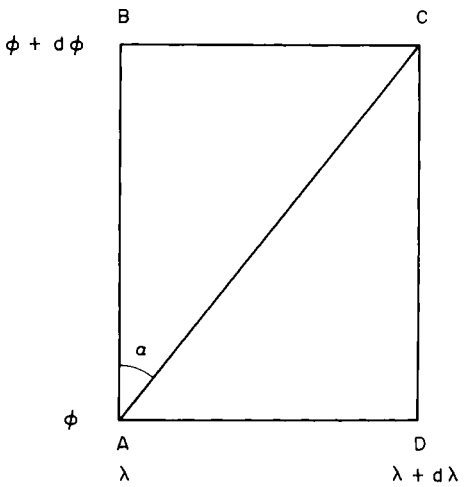


FIG. 5.13 An infinitely small spherical quadrilateral $ABCD$, formed by the intersection of the parallels φ and $\varphi + d\varphi$ with the meridians λ and $\lambda + d\lambda$.

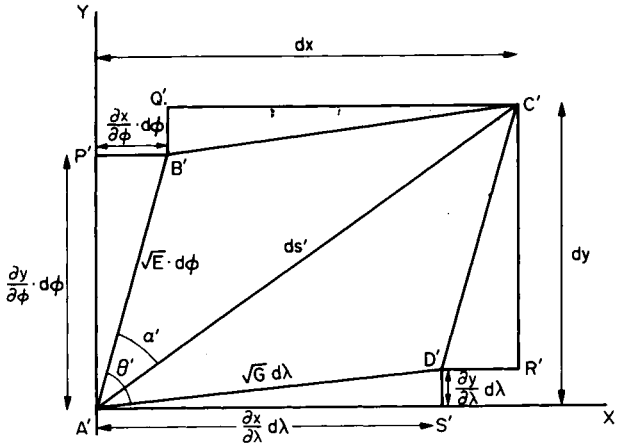


FIG. 5.14 The plane representation $A'B'C'D'$ of the infinitely small spherical quadrilateral illustrated in Fig. 5.13.

tude, x and y as being $d\varphi$, $d\lambda$, dx and dy respectively. This is the usual notation used to indicate infinitesimally small increments. Consequently the four points on the globe are shown in Table 5.02. From equation (3.12) we may express the length of the element of the meridional arc through A as

$$ds_m = R \,.\, d\varphi \tag{5.09}$$

where R is the radius of the globe.

From equation (3.17) the length of the element of the arc of the parallel through A is

$$ds_p = R \,.\, \cos\varphi \,.\, d\lambda \tag{5.10}$$

Moreover, since the angles at the four corners $ABCD$ are right angles, we may use Pythagoras' Theorem to find the length of the diagonal arc element AC. Thus

$$ds^2 = ds_m^2 + ds_p^2 \tag{5.11}$$

TABLE 5.02

Point	Latitude	Longitude
A	φ	λ
B	$\varphi+d\varphi$	λ
C	$\varphi+d\varphi$	$\lambda+d\lambda$
D	φ	$\lambda+d\lambda$

or

$$ds = (R^2 . d\varphi^2 + R^2 . \cos^2 \varphi . d\lambda^2)^{1/2} \qquad (5.12)$$

On the plane, the point $A' = (x, y)$ and $C' = (x+dx, y+dy)$. In order to aid further interpretation, we construct the lines $A'S'$, $B'P'$, $C'Q'$ and $D'R'$ parallel to the x-axis. We also construct the lines $A'P'$, $B'Q'$, $C'R'$ and $D'S'$ parallel to the y-axis. At this stage it is desirable to introduce a word of warning about the understanding of the equations which follow. These are presented in the logical order in which they may be derived, and they make use of the symbolic notation to be expected in mathematical writings. However the rigorous derivation of them calls for a higher standard of mathematical competence than is necessary in most other parts of this book. Consequently we take further short-cuts. We do not attempt to prove these equations algebraically, but merely present the important results. Each of these may be interpreted in geometrical terms, using Fig. 5.14 as the guide, and it is more important for the beginner to understand this part of the argument than the algebraic gymnastics which led to the results. The reader who is anxious for a rigorous mathematical derivation of the theory is referred, for example, to Richardus and Adler (1972). In order to demonstrate how the equations which follow have practical application, we give an example of their use at the end of this chapter.

Each of the lines represented in Fig. 5.14 has a geometrical significance which may be represented symbolically using the notation of *partial differentiation*. Thus $A'B'$ represents the arc of the meridian through A', and $A'D'$ is the arc of the parallel through the same point. $A'C'$ represents any arc through A' which makes the bearing α' with the meridian through A'. The additional construction lines represent the following variables:

- $A'P'$ represents the increment in y which results from an increase $d\varphi$ in latitude. This may be expressed symbolically by the term

$$(\partial y/\partial \varphi)d\varphi$$

- $P'B'$ is the increment in x which results from the same increase $d\varphi$ in latitude. This may be expressed symbolically by the term

$$(\partial x/\partial \varphi)d\varphi$$

- $A'S'$ represents the increment in x resulting from an increase $d\lambda$ in longitude. This may be expressed symbolically by the term

$$(\partial x/\partial \lambda)d\lambda$$

- $D'S'$ is the increment in y resulting from the same increase $d\lambda$ in longitude. This may be expressed symbolically by the term

$$(\partial y/\partial \lambda)d\lambda$$

From Fig. 5.14 we can see that the increment dx between A' and C' is composed of the two linear elements $B'P'$ and $Q'C'$ or

$$\mathrm{dx} = B'P' + Q'C'$$

Substituting the appropriate terms corresponding to these linear elements

$$\mathrm{dx} = (\partial \mathrm{x}/\partial\varphi)\mathrm{d}\varphi + (\partial \mathrm{x}/\partial\lambda)\mathrm{d}\lambda \tag{5.13}$$

In calculus this is known as the *total differential of x*. Similarly we can see that the increment dy between A' and C' is also composed of two linear elements $A'P'$ and $B'Q'$. Thus

$$\begin{aligned}\mathrm{dy} &= A'P' + B'Q' \\ &= (\partial \mathrm{y}/\partial\varphi)\mathrm{d}\varphi + (\partial \mathrm{y}/\partial\lambda)\mathrm{d}\lambda \end{aligned} \tag{5.14}$$

which is the total differential of y. Both of these expressions may be derived algebraically from the functions (5.01) and (5.02). Many elementary textbooks on the calculus demonstrate this.

From the application of Pythagoras' Theorem to the right-angled triangles in Fig. 5.14, the sides and diagonal of the figure $A'B'C'D'$ may be expressed as

$$A'B'^2 = B'P'^2 + A'P'^2 \tag{5.15}$$

$$A'D'^2 = A'S'^2 + D'S'^2 \tag{5.16}$$

$$A'C'^2 = B'P'^2 + Q'C'^2 + A'P'^2 + B'Q'^2 \tag{5.17}$$

$$= \mathrm{dx}^2 + \mathrm{dy}^2 \tag{5.18}$$

Substituting the right-hand sides of equations (5.13) and (5.14), we obtain for the diagonal arc $A'C' = \mathrm{ds}'$, the expression

$$\mathrm{ds}'^2 = [(\partial \mathrm{x}/\partial\varphi)\mathrm{d}\varphi + (\partial \mathrm{x}/\partial\lambda)\mathrm{d}\lambda]^2 + [(\partial \mathrm{y}/\partial\varphi)\mathrm{d}\varphi + (\partial \mathrm{y}/\partial\lambda)\mathrm{d}\lambda]^2 \tag{5.19}$$

Gaussian fundamental quantities

Some simplification of this equation can be obtained if the following expressions are substituted:

$$E = (\partial \mathrm{x}/\partial\varphi)^2 + (\partial \mathrm{y}/\partial\varphi)^2 \tag{5.20}$$

$$F = [(\partial \mathrm{y}/\partial\varphi) . (\partial \mathrm{y}/\partial\lambda)] + [(\partial \mathrm{x}/\partial\varphi) . (\partial \mathrm{x}/\partial\lambda)] \tag{5.21}$$

$$G = (\partial \mathrm{x}/\partial\lambda)^2 + (\partial \mathrm{y}/\partial\lambda)^2 \tag{5.22}$$

leading to the more convenient expression

$$\mathrm{ds}'^2 = E . \mathrm{d}\varphi^2 + 2F . \mathrm{d}\varphi . \mathrm{d}\lambda + G . \mathrm{d}\lambda^2 \tag{5.23}$$

From the equations which have been derived it is now possible to determine three scales which refer to the point A' on the map.

(1) *The scale along the meridian*
This is the ratio

$$A'B'/AB = h$$

Since

$$\begin{aligned} A'B'^2 &= B'P'^2 + A'P'^2 \\ &= [(\partial x/\partial\varphi)d\varphi]^2 + [(\partial y/\partial\varphi)d\varphi]^2 \end{aligned}$$

and from (5.20)

$$A'B' = \sqrt{E}\,.\,d\varphi \tag{5.24}$$

The arc element $AB = ds_m$ has already been determined in equation (5.09). Therefore

$$\begin{aligned} h &= (\sqrt{E}\,.\,d\varphi)/(R\,.\,d\varphi) \\ &= \sqrt{E}/R \end{aligned} \tag{5.25}$$

and since we have to relate this scale to the principal scale, we put $R = 1$ so that

$$h = \sqrt{E} \tag{5.26}$$

(2) *The scale along the parallel*
This is the ratio

$$A'D'/AD = k$$

Since

$$\begin{aligned} A'D'^2 &= A'S'^2 + D'S'^2 \\ &= [(\partial x/\partial\lambda)d\lambda]^2 + [(\partial y/\partial\lambda)d\lambda]^2 \end{aligned}$$

and, from (5.22)

$$A'D' = \sqrt{G}\,.\,d\lambda \tag{5.27}$$

The arc element $AD = ds_p$ has already been found in equation (5.10). Therefore

$$k = (\sqrt{G}\,.\,d\lambda)/(R\,.\cos\varphi\,.\,d\lambda) \tag{5.28}$$

This simplifies to

$$k = \sqrt{G}/(R\,.\cos\varphi) \tag{5.29}$$

or, where $R = 1$,

$$k = \sqrt{G}/\cos\varphi \tag{5.30}$$

(3) *The scale along any arc through A which makes the bearing α with the meridian through A*

This is the general expression illustrated by the ratio A′C′/AC or ds′/ds, which we will denote by μ. The value for $A'C'^2$ has been given in equation (5.23) and that for AC in (5.12). Hence we may write

$$ds'/ds = [E.d\varphi^2 + 2F.d\varphi.d\lambda + G.d\lambda^2]^{1/2}/[R^2.d\varphi^2 + R^2.\cos^2\varphi.d\lambda^2]^{1/2} \tag{5.31}$$

or, putting R = 1, as before

$$\mu = [E.d\varphi^2 + 2F.d\varphi.d\lambda + G.d\lambda]^{1/2}/[d\varphi^2 + \cos^2\varphi.d\lambda^2]^{1/2} \tag{5.32}$$

The scales along the meridian, the parallel or in any direction are known as the *particular scales* at the point and these may now be defined as: * *the relation between an infinitesimal linear distance in any direction at any point on a map projection and the corresponding linear distance on the globe.*

The idea of direction is contained in the angles α' and θ' on the map. We see in Fig. 5.14 that α' is the bearing of the line $A'C'$ measured at A' and corresponds to the bearing $AC = \alpha$ on the globe. The angle θ' is the angle made at A' by the intersection of the meridian and parallel on the map. On the globe this is, of course, a right angle.

It can be shown that

$$\cos\theta' = F/[h.k.\cos\varphi] \tag{5.33}$$

The angle α' may also be shown to be a function of E, F and G so that it is also possible to express (5.32) in terms of α'. This has the form

$$\mu_\alpha^2 = (E/R^2)\cos^2\alpha + (F/R^2\cos\varphi)\sin 2\alpha + (G/R^2\cos^2\varphi)\sin^2\alpha \ldots \tag{5.34}$$

where μ_α is the particular scale in the direction α'. Since E, F, and G change continuously with both latitude and longitude, the particular scales vary with position on the map. Since μ can also be expressed in terms of bearing, it follows that, at any given point, the particular scales also vary with direction about that point.

It follows that any number of particular scales can be evaluated for a point, but, in practice, only four of these are needed for the subsequent analysis of the distortion characteristics of any map projection. These are:

- the particular scale along the meridian, h, from (5.26);
- the particular scale along the parallel, k, from (5.30);
- the maximum particular scale, a, at the point;
- the minimum particular scale, b at the point.

The maximum and minimum particular scales remain to be determined.

CHAPTER 6

The ellipse of distortion

> ... an engineer should use mathematics as a tin-opener is used to open tins of meat. The mathematician also uses mathematics as a tin-opener, but to open tins of tin-openers. Sometimes he is content to indicate the bare existence of a symbolic tin-opener without reference to a tin of anything. He is quite right to do this in pursuit of pure knowledge; and it is our fault if we do not fully appreciate that his objects frequently differ from ours.
>
> M. Hotine, *Empire Survey Review*, 1946

Tissot's Theorem and the principal directions

Most of the foregoing analysis had been undertaken by Gauss in the early years of the nineteenth century. The next major advance in the mathematical theory of map projections was made by N. A. Tissot in the 1850s. He proposed the theorem which bears his name and also developed the concept of the *ellipse of distortion* which is also known as *Tissot's Indicatrix*.

Tissot's Theorem was stated by him as follows:

> *Whatever the system of projection there are, at every point on one of the surfaces and, if angles are not preserved, there are only two of them, such that the directions which correspond to them on the one surface also intersect one another at right angles.*

Tissot's original reasoning is easy to follow. If a point A on the globe represents the intersections of two arcs AB and AC making the angle θ [Fig. 6.01(a)] the corresponding points on the plane are A', B' and C', and the corresponding angle is θ'. We assume that $\theta \neq \theta'$ but that both of them are acute angles. If the line AC is rotated in an anticlockwise direction about A until it is an obtuse angle [Fig. 6.01(b)] it has, at some stage, passed through the angle $\theta = 90°$ during this rotation. Similarly, if the line $A'C'$ is also rotated about A' until it is an obtuse angle, then at some stage $\theta' = 90°$. Where $\theta = \theta' = 90°$, the two orthogonal directions have been defined. These are called the *principal directions*.

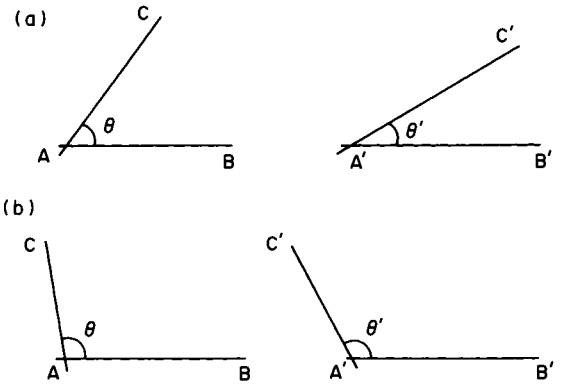


FIG. 6.01 The concept of principal directions. The diagrams on the left relate to the spherical surface and those on the right to its plane representation. In (a), $\theta < 90°$ and $\theta' < 90°$. In (b) $\theta > 90°$ and $\theta' > 90°$. The principal directions are defined where $BAC = B'A'C' = 90°$.

The ellipse of distortion

The next stage in the argument is the most difficult to prove rigorously but simply. The idea is simple enough; namely that an infinitely small circle on the surface of the globe will be transformed on the plane into an infinitely small ellipse whose semi-axes lie along the two principal directions. Reference to Figs 5.01 and 5.02 indicates that the idea is plausible, so we make another massive short-cut and assume it to be proved. The reader who insists upon a proof will find this in a number of advanced textbooks published outside Britain, for example Reignier (1957), Fiala (1957), Richardus and Adler (1972).

Figure 6.02 illustrates a point A on the globe which has geographical coordinates (φ_a, λ_a). AC represents an infinitely short arc, ds, which corresponds to the arc AC in Fig. 5.13 and the preceding section defining particular scales. Since scale is constant on the curved surface of the globe and everywhere equal to the principal scale, the locus of all points such as C traces the circumference of a circle with centre A and radius ds. Since we have set the principal scale $\mu_0 = 1$, it is convenient to make $ds = 1$.

Figure 6.03 illustrates the corresponding figure on the plane. The lines $A'I'$ and $A'II'$ represent the principal directions through the point A' and we use these to define corresponding coordinate axes in both Figs 6.02 and 6.03. In Fig. 6.03 the line $A'C'$ corresponds to the arc element ds′ which, in Fig. 5.14, made the angle α' with the meridian through A'. However, it is now necessary to refer angles to one of the principal directions so we define the angle $IAC = u$ and $I'A'C' = u'$. Because the length of the arc element ds′ varies continuously with direction, or, in

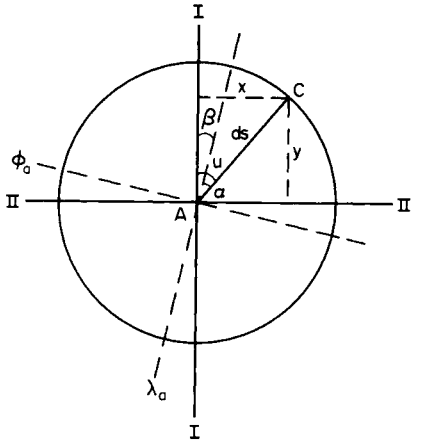


FIG. 6.02 The representation of an infinitely small circle upon the spherical surface.

other words, according to u, it follows that the locus of points such as C' trace the circumference of the ellipse. Let $C = (x, y)$ and $C' = (x', y')$, both systems having the principal directions as axes.

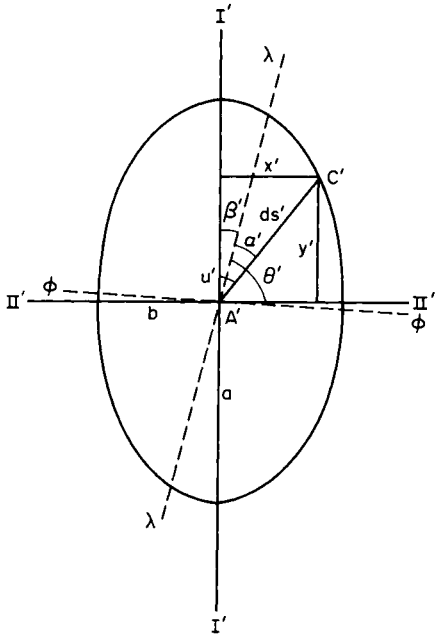


FIG. 6.03 Tissot's Indicatrix or the ellipse of distortion. The deformation of the infinitely small circle illustrated in Fig. 6.02 into an ellipse by the transformation to the plane. Compare also Figs 5.01 and 5.02.

The lengths of the two semi-axes of the ellipse may be expressed as

$$a = y'/y \tag{6.01}$$

$$b = x'/x \tag{6.02}$$

so that

$$x' = b \,.\, x \tag{6.03}$$

$$y' = a \,.\, y \tag{6.04}$$

Moreover

$$x' = ds' \,.\, \sin u' \tag{6.05}$$

$$y' = ds' \,.\, \cos u' \tag{6.06}$$

$$x = ds \,.\, \sin u = \sin u \tag{6.07}$$

$$y = ds \,.\, \cos u = \cos u \tag{6.08}$$

Substituting from equations (6.05)–(6.08) in (6.03) and (6.04)

$$ds' \,.\, \sin u' = b \,.\, \sin u \tag{6.09}$$

$$ds' \,.\, \cos u' = a \,.\, \cos u \tag{6.10}$$

Combination of these equations leads eventually to

$$ds'^2 = a^2 \cos^2 u + b^2 \sin^2 u \tag{6.11}$$

Two of the particular scales at A' refer to the scales along the meridian and parallel. If β is the angle on the globe between the principal direction I and the meridian λ_a through A, with the corresponding angle β' on the map, from (6.11)

$$h^2 = a^2 \cos^2 \beta' + b^2 \sin^2 \beta' \tag{6.12}$$

$$k^2 = a^2 \sin^2 \beta' + b^2 \cos^2 \beta' \tag{6.13}$$

Adding equations (6.12) and (6.13), the terms in β' equal unity (because $\sin^2 \beta' + \cos^2 \beta' = 1$), therefore

$$h^2 + k^2 = a^2 + b^2 \tag{6.14}$$

This is the algebraic expression for the *First Theorem of Apollonius*, well known in plane coordinate geometry, which states that the sum of the squares of two conjugate diameters of an ellipse is constant.

The *Second Theorem of Apollonius* states that the area of the parallelogram formed by two conjugate semi-diameters of an ellipse is equal to the area of the rectangle formed by the semi-axes of that ellipse. In the present notation this may be expressed as

$$h \,.\, k \,.\, \sin \theta' = a \,.\, b \tag{6.15}$$

Equations (6.14) and (6.15) are valuable to the analysis of the distortion characteristics of any map projection, for they permit evaluation of a and b from known values of h, k and θ'. Thus

$$h^2 \pm 2h \,.\, k \sin\theta' + k^2 = a^2 \pm 2a \,.\, b + b^2 \tag{6.16}$$

whence

$$a \pm b = (h^2 + k^2 \pm 2h \,.\, k \,.\, \sin\theta')^{1/2} \tag{6.17}$$

Area scale

The area of a small quadrilateral, such as $A'B'C'D'$ in Fig. 5.14, may be defined as $A'B' \,.\, A'D' \,.\, \sin\theta'$. Thus

$$p = h \,.\, k \,.\, \sin\theta' \tag{6.18}$$

which is the left-hand side of (6.15). Consequently we may also write

$$p = a \,.\, b \tag{6.19}$$

The parameter p is defined in the same units as the particular scales; therefore it is known as the *area scale*.

Angular deformation

From the difference between the angles u and u', both being referred to the same principal direction, it is possible to evaluate the alteration in direction of the line $A'C'$ as follows:

$$\tan u' = (b/a) \tan u \tag{6.20}$$

and

$$\tan u \pm \tan u' = \tan u \pm (b/a) \tan u \tag{6.21}$$

It can be shown that

$$[\sin(u - u')/\cos u \,.\, \cos u'] = [(a + b)/a] \tan u \tag{6.22}$$

and

$$[\sin(u - u')]/[\cos u \,.\, \cos u'] = [(a - b)/a] \tan u \tag{6.23}$$

Dividing (6.23) by (6.22)

$$\sin(u - u') = [(a - b)/(a + b)] \sin(u + u') \tag{6.24}$$

This equation will have the maximum value when $\sin(u + u') = 1$, corresponding to $(u' + u) = 90^\circ$. There will be four such directions located one in each of the four quadrants of the coordinate axes defined by the principal directions. If an angle is composed of two such directions, so

that each side of the angle has been deflected through the maximum amount, we have a quantity called the *maximum angular deformation*, ω, at the point. It follows from (6.24) that

$$\sin(\omega/2) = (a-b)/(a+b) \tag{6.25}$$

and this is the formula which is most commonly used to find this parameter for any point in a map projection.

Summary of the main conclusions derived in Chapters 5 and 6

These chapters have contained some fairly difficult mathematical ideas and unfamiliar concepts. To help the beginner to keep track of the argument it is worth summarising the main conclusions which have so far been obtained.

1. It is necessary to distinguish two kinds of scale on any map projection.
2. The principal scale, μ_0, is the nominal scale of the map. It can only be preserved at all points and in all directions on the curved surface of the globe. On the map the principal scale can only be preserved at certain points or along certain lines.
3. These are known as points or lines of zero distortion.
4. The principal scale is allocated a numerical value of 1·0.
5. The particular scales, μ, at any point on a map projection are those defined for infinitely short arcs in different directions. These are expressed as a decimal fraction or multiple of μ_0.
6. Particular scales vary throughout the map according to position and direction.
7. Two particular scales through any point can always be determined. These are the particular scale along the meridian, h and that along the parallel, k.
8. Tissot's Theorem demonstrates that at every point there are two orthogonal principal directions which are perpendicular to one another on both the globe and map.
9. An infinitely small circle on the globe will be represented on the map by an infinitely small ellipse, known as Tissot's Indicatrix or the ellipse of distortion.
10. The axes of the ellipse of distortion correspond to the two principal directions and the maximum and minimum particular scales, a and b at this point occur in these directions.
11. These particular scales may be evaluated if h and k are known, together with the angle θ' made by the intersection of the meridian and parallel on the map at this point.
12. From the Second Theorem of Apollonius it is possible to derive the

area scale (or exaggeration of area), p, which relates the areas of infinitely small figures on the globe to the corresponding figures on the map.

13. It is also possible to evaluate the maximum angular deformation, ω, from the maximum and minimum particular scales at a point.

The special properties of map projections

Despite the fact that the principal scale can only be preserved along certain lines or at certain points in a projection; despite the fact that the particular scales are variables in both position and direction on the map, it is possible to create certain special combinations of particular scale which may be maintained through a map projection of the whole world, excepting only at the singular points where Tissot's theory does not apply. These arrangements of the particular scales may be called the *special properties of a map projection* (some writers call them *the* properties) which may be defined as *the properties of a projection which arise from the mutual relationship between the maximum and minimum particular scales at any point and which are preserved at all except the singular points of a map.*

The present author (Maling, 1968b) has suggested that there are about a dozen different arrangements of the particular scales which may be regarded as special properties, but only four of these are really important. These are the properties of:

- conformality,
- equivalence,
- equidistance,
- minimum-error representation.

Conformality

A conformal map is one in which

$$a = b \tag{6.26}$$

at all points on the map. It follows that, if this condition can be satisfied, the infinitely small circle on the surface of the globe will always project as a circle on the plane. Moreover, since the maximum angular deformation is determined from the relationship $(a-b)/(a+b)$ in equation (5.59) it follows that where a and b are equal, $\omega = 0°$. Thus a conformal map projection has no angular deformation, or, to paraphrase part of Tissot's Theorem quoted on p. 100, *angles are preserved.* This is the essential and important special property of all conformal projections. It is an essential requirement for any map which is to be used for measure-

ment of angles. Hence conformal projections are used as the bases for navigation charts, topographical maps and military maps.

The fact that an infinitesimally small circle on the globe remains a circle on the map implies a further property of a conformal map, namely that the *shapes* of objects are also preserved. However, this statement must be accepted only with certain reservations. The condition expressed in equation (6.26) is not equivalent to the statement that $a = b = 1$. Conformality can be obtained only at the expense of increasing particular scales by the same amount in all directions. This means that the area scale increases according to the square of the particular scale. The result is that a circle on a point of zero distortion remains a circle near the edge of the map, but the size of it has increased considerably. Hence we uncover the paradox that although a conformal map provides a good representation of shapes for a small area round every point, the rapid increase in the particular scales away from the points or lines of zero distortion make these projections less suitable for representing the shapes of large terrestrial features like continents and oceans.

The alternative name for this property is *orthomorphism*, but the use of this term has tended to divert attention from correct angular representation to the much less important consideration of shape. In mathematics a conformal transformation is one in which every angle retains its original size. This is precisely what we mean by (6.26); therefore we prefer to use the adjective conformal rather than orthomorphic. However, the reader is referred to the correspondence, Arden-Close (1944), Hotine (1945–1946), Lee (1946), Lenox-Conyngham (1944) for different opinions concerning this usage.

It follows that, in any projection for which $\omega = 0°$, all graticule intersections are orthogonal. This must be true irrespective of the nature of the mapped parallels and meridians which are often complicated curves. If a conformal projection is composed of curved parallels and meridians it is necessary to imagine two tangents, one to each line, drawn at the graticule intersection. These two tangents are perpendicular to one another. The converse does not necessarily apply. Thus a map projection in which the parallels and meridians all intersect at right angles is not necessarily a conformal projection.

Equivalence

An equal-area map is one in which

$$a.b = 1 \tag{6.27}$$

It therefore follows that

$$a = 1/b \tag{6.28}$$

$$b = 1/a \tag{6.29}$$

or the maximum and minimum particular scales are reciprocals of one another. It follows that although the ellipses of distortion may have considerable ellipticity, they have uniform area. Moreover, the principle of equivalence may also be maintained for areas of finite size and an important aspect in the derivation of equal-area map projections of different classes has been the ability to argue that the whole or part of the generating globe is mapped into a square, rectangle, circle, ellipse or other geometrical figure having the same area as the required part of the globe.

The equal-area map projections are most important in the field of distribution mapping of statistical variables. For example, if it is required to map population, agricultural or industrial statistics, this may be done by plotting many symbols, such as dots, each representing a particular number or quantity of the variate. An important aspect of interpretation of such a map is the visual impression of *density* of population, agricultural production or industrial output as this varies from place to place in a country or continent. This visual impression is, of course, created by the concentration of many such dots in some places contrasted with sparser distribution of them elsewhere. If the base map upon which such distributions are plotted is truly equal-area, the visual impression is likely to be correct. If, however, the map is not equal-area, the visual impression of density is upset by the wholly artificial crowding or dispersion of symbols. We may also wish to measure the area occupied by some distribution, such as a category of land use, on a small-scale map. Then it is desirable to use a map in which there is no exaggeration of area. See Maling (1989) for an analysis of this problem, and some of the ways of overcoming it when the ideal map is not available.

Equidistance

The third important mathematical property which may be satisfied is that one particular scale is made equal to the principal scale throughout the map. Usually this is the meridional scale so that for equidistance we may write

$$h = 1{\cdot}0 \tag{6.30}$$

thereby creating a projection in which all the parallels intersect all the meridians at a separation corresponding to the arc distance between the parallels on the globe. The alternative is to make $k = 1$ throughout the map. This property arises incidentally in the derivation of certain map projections, but it is less valuable than preserving the principal scale along great circle arcs.

Since we have specified that one particular scale is equal to unity it follows that equidistance is incompatible with both conformality and equivalence. Clearly if $a = 1$ in an equidistant projection, the conditions specified by either (6.26) or (6.27) would lead us once more to the perfect but impossible solution.

Equidistance is a less valuable property than either conformality or equivalence because it is seldom desirable to have a map in which distances may be measured correctly in only one direction. However, an equidistance map is a useful compromise between the two extremes represented by conformal and equal-area maps. Thus the area-scale of an equidistant map increases more slowly than that of a conformal map. The maximum angular deformation of an equidistant map increases more slowly than that of an equal-area map. Consequently equidistant map projections are often used in atlas maps, strategic planning maps and similar representations of large parts of the earth's surface in which it is not essential to preserve either of the other properties.

Minimum-error representation

We have seen that the three special properties which have been described are mutually exclusive of one another. Minimum-error representation is a rather different kind of property because it may be combined with some other special property. For example a *minimum-error conformal projection* of a particular class may be specified for a particular purpose. However minimum-error representation can also be considered to be a special property in its own right, giving rise to what may be termed an *absolute minimum-error projection.* The idea is well described by the older term *balance of errors* used by Airy to describe the minimum-error projection associated with his name. We already know that a and b are the maximum and minimum particular scales at any point. Since we specify that the principal scale is equal to unity, it follows that the *scale errors* along the principal directions through a point may be expressed respectively as

$$e_1 = 1 - a \tag{6.31}$$

$$e_2 = 1 - b \tag{6.32}$$

The idea implicit in any minimum-error map projection is to balance these errors so that *the sums of the squares of the scale errors throughout the map as a whole are a minimum value.* For example it is necessary to find expressions for (r, θ) which satisfy the condition that

$$\int_{z=0}^{z=\beta} [(1-a)^2 + (1-b^2)] \,.\, \sin z \,.\, dz = \text{minimum} \tag{6.33}$$

It is necessary to specify the limits of the area to be mapped in which

these conditions must be satisfied. Thus (5.67) must be expressed as the definite integral which indicates the size of the area to be mapped. In (5.67) we have taken the simplest case of a map with a circular boundary and point of zero distortion at the centre of this circle. We shall recognise that later as an *azimuthal projection.* Then the definite integral indicates summation of the sums of the squares of the scale errors at all points from the centre of the map (where $z = 0$) to the edge of the map (where $z = \beta$). Clearly the expression of the minimum-error conditions for many projections is algebraically quite difficult to follow. There have been two important works on the subject published in the past 70 years. The classic work is that of Young (1920); the contemporary study is that by Snyder (1985).

The practical use and interpretation of the distortion characteristics of a map projection

In this chapter we have derived a series of algebraic expressions for the four important particular scales at any point. The additional parameters, p and ω may be derived from these particular scales and the special properties of any map projection are also defined in terms of the particular scales and distortion parameters.

It is now desirable to show what value these characteristics have in helping us to describe a particular map projection. Even more important, they give us some clues about a logical and systematic way of choosing which map projection is suitable for a particular purpose.

Tabular presentation of distortion characteristics

Usually the values for the particular scales are calculated for a fairly widely spaced graticule, for example, 10° or 15° of latitude and longitude for a world or hemispheric map. There is no reason why the information should not be calculated for every 1°, or for that matter every 1′ or 1″ apart from the sheet volume of the output. On the other hand, if it is only done for every 20° or 30° some salient features of the given projection may be missed. The results of the computations may be listed in a form such as is given in Table 6.01. This projection is illustrated by Fig. 6.04. Although we have not yet defined what we mean by a cylindrical projection, it can be seen that the world map is represented by a rectangular outline, and both the parallels and meridians are families of parallel straight lines. We may conduct interpretation of Table 6.01 in the following fashion:

(1) *We look for evidence of the location of the lines or points of zero distortion.* Since the principal scale is conventionally expressed as $\mu_0 = 1{\cdot}0$, we look for values corresponding to this in the columns for the

TABLE 6.01 *Particular scales and distortion characteristics for the Cylindrical equal-area projection (Lambert)*

Latitude	Particular scales		Area scale	Maximum angular deformation
φ	$k = a$	$h = b$	p	ω
0°	1·0000	1·0000	1·0000	0°
15°	1·0335	0·9659	1·0000	3° 58′
30°	1·1547	0·8660	1·0000	16° 25′
45°	1·4142	0·7071	1·0000	38° 57′
60°	2·0000	0·5000	1·0000	73° 44′
75°	3·8637	0·2588	1·0000	121° 57′
90°	∞	0·0000	–	180°

particular scales. We find that both scales are equal to unity in the first line corresponding to $\varphi = 0°$. We also note that $p = 1{\cdot}0000$ and $\omega = 0°$. This confirms that the principal scale is preserved along the equator, which is therefore a line of zero distortion. We can also see that the particular scales do not equal unity elsewhere. Consequently the equator is the only line of zero distortion.

(2) *We look for evidence about special properties.* This must be a relationship which is established for the whole projection. From the preceding section it is likely to be of the form $a = b$, $a = 1/b$, $h = 1{\cdot}0000$ or $k = 1{\cdot}0000$. A conformal map will have $\omega = 0°$ throughout, and an equal-area map will have $p = 1{\cdot}0000$ throughout. We find the evidence that this is an equal-area projection from the constant value for p in column 4 of Table 6.01. It may be argued that the use of the words *equal-area* or *conformal* in the name of the projection should be sufficient evidence about the special properties of it. However, this is not necessarily so. Some projections are commonly only referred to by personal names or titles (Mercator's projection, Bonne's projection or the Twilight projection) which convey none of this information. Sometimes they are incorrectly labelled with an adjective which does not strictly apply to them.

(3) *We look for evidence concerning the principal directions.* In this particular example the parallels and meridians form an orthogonal network and therefore the principal directions coincide with the graticule. Thus $k = a$ and $h = b$. It follows that a projection of this kind is much easier to study than one having principal directions which do not coincide with the parallels and meridians.

(4) *We look for evidence for singular points, characterized by particular scales equal to zero or infinity.* This is shown in the last line of Table 6.01 where $\varphi = 90°$. Here $a = \infty$, $b = 0{\cdot}0000$, p is indeterminate and the maximum angular deformation $\omega = 180°$. All these clues lead us to suppose that the one-to-one correspondence of points does not apply at

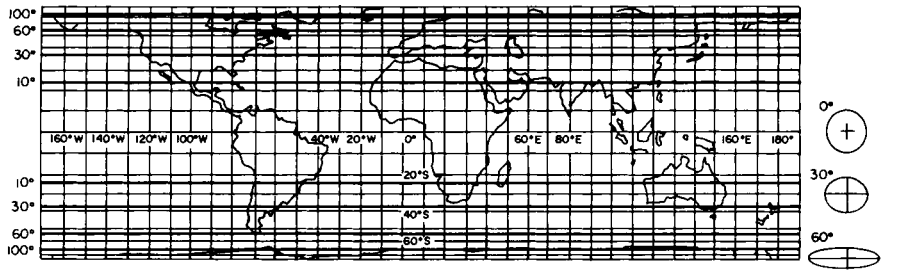


FIG. 6.04 World map based upon the normal aspect of the Cylindrical equal-area projection (Lambert) (No. 1 in Appendix I) showing conventional representation of Tissot's Indicatrix for the parallels 0°, 30° and 60° (on right) and also showing isograms for maximum angular deformation (ω) for 10°, 30°, 60° and 100° (labelling on left of map).

the geographical poles. This is confirmed in Fig. 6.04 by the representation of the poles by means of lines which are the same length as the equator.

(5) *We may study the variations in particular scale with latitude.*

Graphic presentation of deformation

This is done quite simply by plotting graphs for a and b against φ, as shown in Fig. 6.05. Each of the numerical values in the table have been determined for the points at which the parallels intersect a meridian and, in theory, these values relate to the axes of the infinitely small ellipse located at each intersection. If the map is a continuous representation of the spherical surface, as in the present example, and there are no gaps or interruptions, such as are illustrated by Fig. 5.01, we are justified in making the interpretation that particular scales vary continuously and regularly between the points which have been plotted. For example, if $k = 1{\cdot}4142$ in latitude 45° and $k = 2{\cdot}0000$ in latitude 60°, we may interpolate from the graph and approximate value $k = 1{\cdot}55$ for latitude 50°. This may be done with greater accuracy by interpolation within the table, provided that one of the standard methods of numerical interpolation is applied. Simple graphs showing the particular scales plotted against latitude are very useful in assessing the relative merits of several different map projections which might be chosen for a particular job. The gradient and nature of each curve compared with others gives a useful visual appreciation about which of several projections provides least distortion in a particular part of a map. The same kinds of graphs can also be drawn for variations in p and ω. We shall make use of this means of comparison in Chapters 11 and 12.

(6) *We may also use spatial representation of the ellipses of distortion.* Thus, if we plot a and b to some arbitrary but convenient scale we may

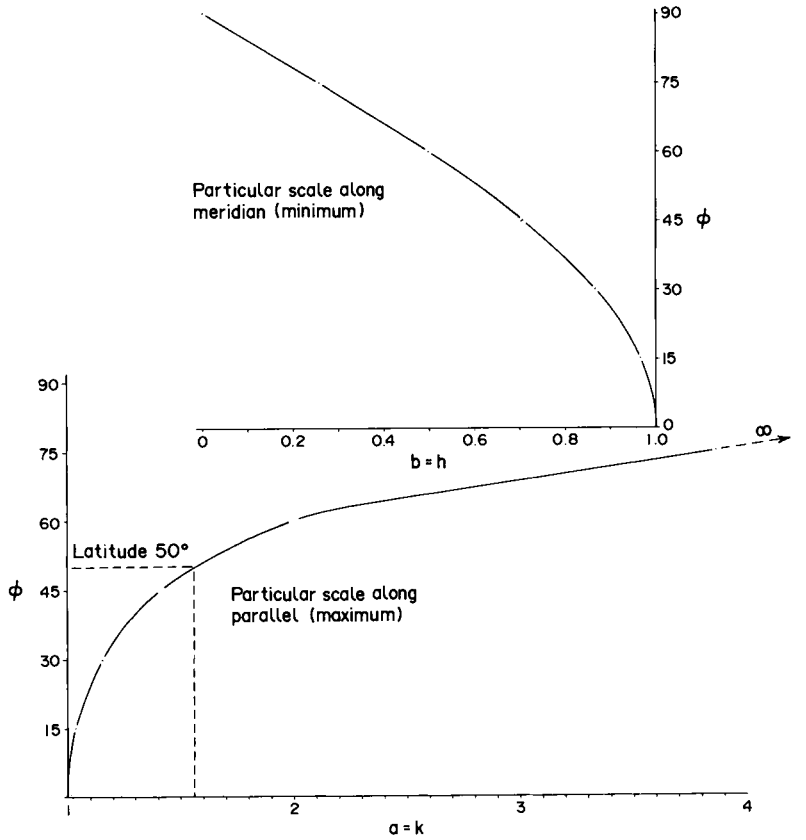


FIG. 6.05 Graphs showing the variations in particular scales with latitude for the Cylindrical equal-area projection, illustrated by Fig. 6.04. These graphs have been plotted from the numerical data for the particular scales in Table 6.01.

construct the ellipses corresponding to different points on the projection. These diagrams provide a generalized picture of deformation from place to place, as illustrated on the right-hand side of Fig. 6.04. Several points about their interpretation should be emphasized. The first is that on the equator the ellipse of distortion is a circle of radius 1·0 units on the arbitrary scale which has been chosen to draw these figures. This, again, confirms that the equator is a line of zero distortion. Secondly, the flattening of the ellipses varies exceedingly, but all of them appear to be of similar size. This is confirmed by the fact that we are dealing with an equal-area projection, so that the areas of the ellipses ought to be the same.

(7) *We may plot a series of isograms indicating constant values for any single parameter.* In this example the variable selected for illustration by

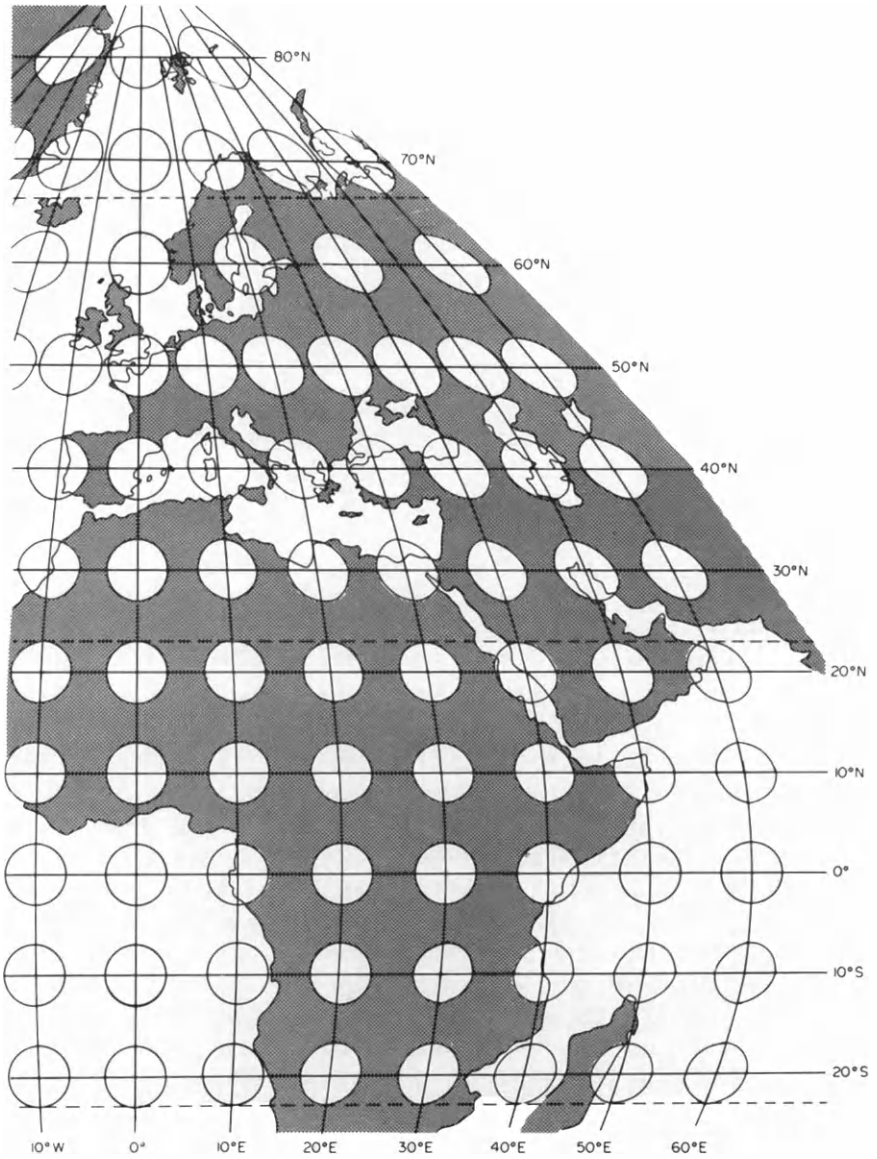
this means is the maximum angular deformation. By determining the latitudes for which $\omega = 10°$, $30°$ etc. we may plot curves (or in this example, straight lines corresponding to parallels of latitude). The pattern of the isograms, which may be improved visually by using variable shading as in many illustrations in this book, give a two-dimensional picture of how distortion varies from place to place, rather than the one-dimensional picture provided by a single graph. This is important in the study of many projections.

Where the particular scales vary with both latitude and longitude, information such as that recorded in Table 6.01 would refer to only one meridian. Thus a table for a 15° world graticule might require up to 338 separate entries for each of the variables a, b, p and ω*. This kind of table is difficult to comprehend, and graphical representation of the variables is practically essential. It can be done by showing ellipses at every graticule intersection as illustrated by Fig. 6.06, but this is an extremely laborious way of doing it unless a digital solution is sought. If such a figure has to be drawn by hand, the dimensions and orientation of each ellipse has to be calculated, plotted and fair-drawn. In Fig. 6.06, although the ellipses are all of the same size and there is a line of identical circles along the equator and central meridian, the shape and orientation of all the other ellipses differ at every graticule intersection. At the time when the first edition of this book was being prepared, the preparation of Fig. 6.06 caused an immense amount of trouble; sufficient to deter us from ever trying to produce another by hand. At about the same time,

*The actual number of entries depends upon the symmetry of the projection about certain parallels and meridians. Some, like the cylindrical projections, are symmetrical about the equator so that the tabulated values are valid for both hemispheres. A projection which is symmetrical about both the equator and a central meridian only requires tabulated values of the particular scales for 79 graticule intersections. See also p. 138.

FIG. 6.06 Part of the Sinusoidal projection (No. 30 in Appendix I) showing a diagrammatic representation of ellipses of distortion at each graticule intersection. This is an equal-area member of the pseudocylindrical class of projections in which the meridians are sine curves. The parallels are equally spaced along the central meridian. Note the following features of these ellipses: (1) that the ellipses along the equator and the Greenwich Meridian are circles, indicating that these are lines of zero distortion; (2) that all the ellipses have the same area, indicating that this is an equal-area projection; (3) that there is an increasing flattening of the ellipses towards the north-eastern part of the map; (4) that the axes of the ellipses do not correspond to the directions of the meridians and parallels, and that the divergence in orientation increases towards the north-eastern edge of the map. This is also confirmed by the increasing obliquity of intersection of the graticule there. Obviously the principal directions, which are the axes of the ellipses, cannot correspond to the graticule. Compare this means of representation with Fig. 7.04(a), p. 132, where isograms for maximum angular deformation are shown.

however, Richardus and Adler (1972) were obtaining graph-plotter output of examples of the same technique used, in their work, to illustrate the deformations of certain conical projections. Indeed it is the only method which they illustrate. Similarly, Snyder and Voxland (1989) use this method to the exclusion of all others.



More commonly the distortion patterns are shown by means of isograms and shading. Figure 6.07 illustrates such a technique applied to a world map in which the isograms do not coincide with the graticule. An important advantage in using these parameters to assess the distortion characteristics and relative merits of a map projection is that the parameters have already been computed for the majority of useful map projections. For example, Reignier (1957) gives tables for most of the better-known projections.

Some other views of Tissot's work

Despite the evident advantages of Tissot's method of describing the distortions which arise in the process of representing one surface upon another, it is important to appreciate that this method has had its critics in the past. Some writers have maintained that a method of evaluation which is derived from the particular scales, and therefore upon infinitesimal areas, is unrealistic. Thus Hinks (1912, 1921a,b) was critical of Tissot's methods and did not attempt to use them. This is the main reason why the methods outlined in this chapter are still seldom described in English works on map projections, whereas they are commonplace in every other European language. Tobler (1964) has also made certain reservations about the validity of interpreting the distortion characteristics of map projections solely in terms of the ellipse of distortion. But critics of the method have tended to ignore the principle outlined in (5) above, that if x and y are continuous functions of φ and λ, the particular scales and derived parameters also increase or decrease continuously and can therefore by mapped. Tobler's published alternative method, which involves the determination of finite errors in computed triangles of different sizes in different parts of a map, is a more elaborate procedure which, at the time of publication, could be tackled only by using a mainframe computer. Moreover, the presentation of the results is tabular and statistical, so that it is difficult to appreciate how distortion can vary from one part of a projection to another. The reader who can obtain access to the very interesting *Atlas for the Selection of Map Projections*, by Ginzburg and Salmanova (1957), will appreciate that simple graphics based upon the six variables which have been defined here can be enormously helpful in deciding which projection is going to be the most useful to serve as the framework of a new map. After all, this is the chief practical reason for wishing to know about the spatial distribution of distortion in a projection. The work by Synder and Voxland (1989), entitled *An Album of Map Projections*, is similarly a most useful graphic guide to the appearance of world projections. However, its practical value is somewhat reduced by only using plots of the ellipse of distortion to illustrate how

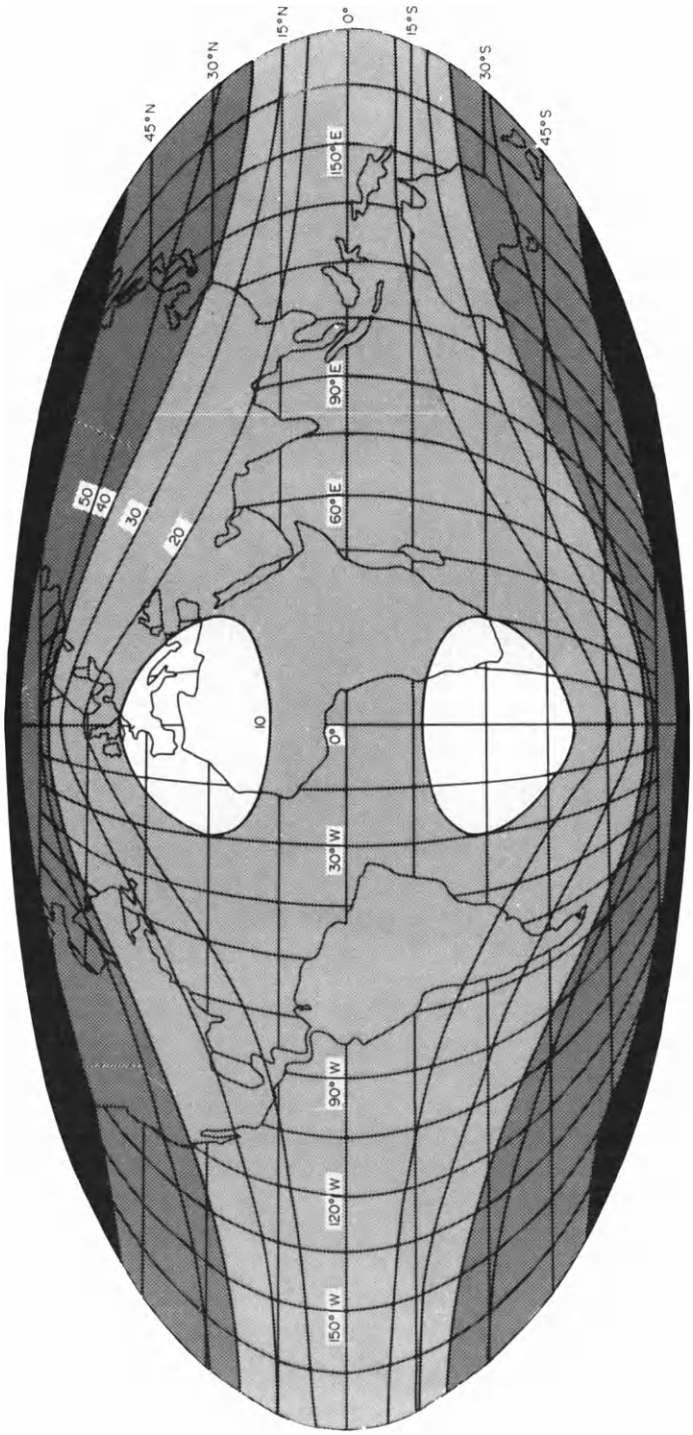


FIG. 6.07 Mollweide's projection for the whole world showing isograms for maximum angular deformation, (ω) at 10°, 20°, 30°, 40°, 50°. Parts of the world where $\omega > 80°$ are shown black. This is an equal-area projection and therefore $p = 1{\cdot}0$ throughout.

deformation changes from place to place on the map. We return to this important subject in Chapters 11 and 12, where the principles of selecting a projection are considered in detail.

Theoretically it is also possible to apply the variations in particular scale as corrections to measurements of distance, angle and area made from maps. But the present author must confess that he has never met anyone from outside Russia who admitted to ever having done this.

Worked example using the equations in Chapters 5 and 6

After such a lengthy algebraic introduction to the theory of distortion it is desirable to show how the variables may be computed to find numerical values for the particular scales and distortion characteristics at a specific point in a projection. The example given here is for a point on the *Hammer–Aitoff* projection (Fig. 6.08) in latitude $\varphi = 60°\text{N}$, longitude $\lambda = 60°\text{E}$. This example has been chosen because both the meridians and parallels are curved and do not intersect at right angles. Consequently no simplification is possible such as occurs when the principal directions coincide with the graticule. Therefore it is necessary to start by finding the numerical values for E, F and G. The formulae which follow are from Maling (1962).

The coordinates for a point on the Hammer–Aitoff projection may be written in the form

$$\text{x} = 2\sqrt{2}\{(\cos\varphi \,.\, \sin \tfrac{1}{2}\lambda)/[1+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda]^{1/2}\} \tag{6.34}$$

$$\text{y} = (\sqrt{2} \,.\, \sin\varphi)/[1+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda]^{1/2} \tag{6.35}$$

The first requirement is to differentiate equations (6.34) and (6.35) with respect to φ and λ. This is by far the most difficult stage in the solution so we do not expect the beginner to understand the derivation of the four following equations

$$\partial\text{x}/\partial\varphi = -\sqrt{2}\{[\sin\varphi \,.\, \sin \tfrac{1}{2}\lambda(2+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)]/[1+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda]^{3/2}\} \tag{6.36}$$

$$\partial\text{y}/\partial\varphi = [\cos\varphi(2+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)+\cos \tfrac{1}{2}\lambda]/[\sqrt{2}(1+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)^{3/2}] \tag{6.37}$$

$$\partial\text{x}/\partial\lambda = [\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda(2+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)+\cos^2\varphi]/[\sqrt{2}(1+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)^{3/2}] \tag{6.38}$$

$$\partial\text{y}/\partial\lambda = [1/2^{3/2}] \,.\, [\sin\varphi \,.\, \cos\varphi \,.\, \sin \tfrac{1}{2}\lambda]/[(1+\cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)^{3/2}] \tag{6.39}$$

Once these equations are available, the numerical solution is not difficult using a pocket calculator with hard-wired trigonometric functions, but it

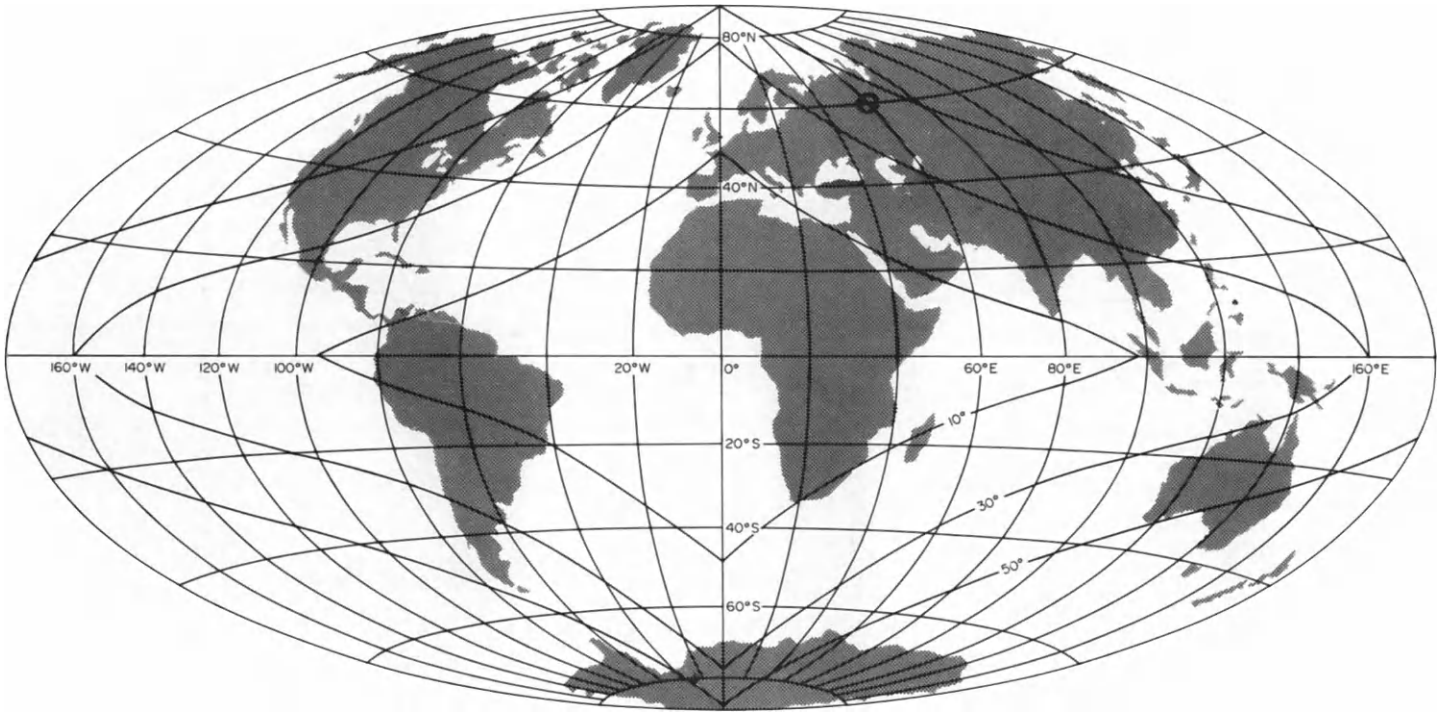


FIG. 6.08 The Hammer–Aitoff projection for the world showing isograms for maximum angular deformation (ω) at 10°, 30° and 50°. The point in latitude 60°N, longitude 60°E, which is indicated by a circle, is the point for which the particular scales and distortion characteristics have been calculated on pp. 120–121. The Hammer–Aitoff projection (No. 37 in Appendix I) is an equal-area member of the polyconic group of projections.

is obviously even easier to write a program to solve them by microcomputer. Note that the term $(1+\cos\varphi \,.\, \cos\frac{1}{2}\lambda)$ appears in three of the equations and the denominator $(1+\cos\varphi \,.\, \cos\frac{1}{2}\lambda)^{3/2}$ occurs in all four. These terms only have to be calculated once for each graticule intersection.

Substituting for $\varphi = 60°$, $\lambda = 60°$ in equations (6.36)–(6.39) gives the following numerical values:

$$\partial x/\partial\varphi = -0{\cdot}8685$$
$$\partial y/\partial\varphi = 0{\cdot}8584$$
$$\partial x/\partial\lambda = 0{\cdot}5373$$
$$\partial y/\partial\lambda = 0{\cdot}0446$$

Then, from (5.20)

$$E = -0{\cdot}8685+0{\cdot}8584 = 1{\cdot}4911$$

and, from (5.25), the particular scale along the meridian is

$$h = \sqrt{E} = 1{\cdot}2211$$

Similarly, from (5.22)

$$G = 0{\cdot}5373^2+0{\cdot}0446^2 = 0{\cdot}2907$$

and from (5.30) the particular scale along the parallel is

$$k = \sqrt{G}/\cos\varphi = 0{\cdot}5391/0{\cdot}5 = 1{\cdot}0783$$

We obtain the third fundamental quantity, F from (5.21)

$$F = -(0{\cdot}8584\times 0{\cdot}0446)-(-0{\cdot}8685\times 0{\cdot}5373) = -0{\cdot}4284$$

From (5.33)

$$\cos\theta' = F/(h\,.\,k\,.\cos\varphi)$$
$$= -0{\cdot}4284/0{\cdot}6584$$
$$= -0{\cdot}6507$$

it follows that $\sin\theta' = 0{\cdot}7594$. From (6.17)

$$(a+b) = h+k+2h\,.\,k\sin\theta'$$
$$= 4{\cdot}6536$$
$$(a+b) = 2{\cdot}1572$$

Similarly

$$(a-b) = h+k-2h\,.\,k\sin\theta'$$
$$= 0{\cdot}6540$$
$$(a-b) = 0{\cdot}8087$$

Therefore

$$a+b = 2{\cdot}1572$$
$$a-b = 0{\cdot}8087$$
$$\overline{\qquad\qquad}$$
$$2a = 2{\cdot}9659$$
$$a = 1{\cdot}4830$$
$$b = 0{\cdot}6743$$

It follows that $a\,.\,b = 0{\cdot}99999$, indicating a small rounding error, but is close enough to 1·0 to confirm that the projection is equal-area. Finally

$$\sin\omega/2 = 0{\cdot}8087/2{\cdot}1572$$
$$\omega/2 = 22^\circ\,01'$$

Therefore

$$\omega = 44^\circ\,02'$$

We therefore obtain the following numerical values for the point 60°N, 60°E:

$$h = 1{\cdot}2211 \qquad k = 1{\cdot}0783$$
$$a = 1{\cdot}4836 \qquad b = 0{\cdot}6743$$
$$p = 0{\cdot}99999 \qquad \omega = 44^\circ\,02'$$

In order to draw satisfactory isograms for ω, it would be necessary to derive such values for at least 50 points on the map.

CHAPTER 7

The appearance, classification and naming of map projections

> Viewed in this light the projections of M. Tissot assume a new aspect, and it is clearly necessary to study them anew, and to master his rather repellent terminology, that seems so superfluously different from that of his compatriot Germain.
>
> A. R. Hinks, *Geographical Journal*, 1921

Introduction

Examination of the illustrations of different map projections which appear in this book indicates the great variety in the shape and detailed appearance of them. Some of the world maps are rectangular in outline, others are bounded by ellipses or more complicated curves. Some projections have rectilinear parallels or meridians; others have various combinations of curved graticule lines. In this chapter we introduce some of the terms which are commonly used to describe the *appearance* of map projections. These may be used in conjunction with distortion theory to *select* and *describe* suitable map projections for particular purposes, or to *recognise* the projection used for a particular map.

If the cartographer has not done his job properly, and has failed to indicate this information, or has described the projection in unfamiliar terms, the critical user has to make a reasoned guess about what projection has been used. The cartographer can communicate with the map user if both understand the same technical terms, but confusion and misinterpretation result if they do not. The subject of map projections is embarrassingly rich in words which mean the same thing. Therefore the beginner who is already struggling to understand many new concepts is also confronted with and confused by duplicate terms. Some of these are synonymous, such as the words 'autogonal' and 'orthomorphic' to mean *conformal*, or the use of 'authalic' or 'orthembadic' instead of *equal-area*. Only two of the six words are necessary.* On the other hand, there are

*Where alternative words are given in this and subsequent chapters, the preferred term is given in italics, and the others are placed in quotation marks.

occasions when different words are needed to make fine but important distinctions. The difference between the definitions for *azimuth* and *bearing* given in Chapter 3, pp. 53–55 illustrates the need for more than one word to describe angles on the spherical surface and the plane.

Modern work on terminology

Nowadays this richness of terminology ought to create fewer problems than it did. In 1964 the International Cartographic Association established a Commission to study the standardisation of technical terms. This led in turn to the creation of a British Working Group of Terminology and publication by the Royal Society of the *Glossary of Technical Terms in Cartography* (Royal Society, 1966). Similar work was in progress in other countries, and the culmination of all this work was publication of the *Multilingual Dictionary of Technical Terms in Cartography* (ICA, 1973). The author assisted the UK Working Group in their deliberations about map projections, and published a specialised multilingual glossary of usage in the study of map projections in Maling (1968b) much of which, in turn, was incorporated into ICA (1973). All these works indicated the preferred usage for future English contributions to the subject, and these words are used throughout the present book. Notwithstanding this work, which has now been available for more than quarter of a century, we still find anomalous usage. For example, in an otherwise first-rate introduction to the subject, the Open University television programme, *M203: Maps*, which was made in 1978, two common map projections are described with names which were evidently known only to the producer of that programme, so that the OU mathematics student learns two names which are unknown in cartography and which are not to be found in any atlas. The Cylindrical equal-area projection is renamed 'Lambert horizontal' and the Azimuthal equidistant projection is renamed 'the great circle map'. Similarly, the Royal Geographical Society, which really ought to have known better, have recently (RGS, 1989) referred to the projection formerly used in their logo as 'an upright projection by Sir Henry James'. The projection attributed to Sir Henry James is well enough known and correctly described, but the interested reader can search in vain, in the terminological literature, for a description of the 'upright' version. Perhaps it is the opposite to a 'horizontal' projection.

In order to employ a satisfactory and succinct terminology we must also create some sort of classification system. The total number of map projections which can be described is infinitely great. From this population about 400 projections have been described, though less than one-quarter of them have been named and used. In order to distinguish between them it is desirable to group together those map projections which possess similar attributes, or have related characteristics, into some

kind of ordered system. The student of the subject can visualise how each projection is related to others; to appreciate where each belongs within this vast collection of slightly different kinds of transformation. Moreover, a series of classification terms is helpful in providing each map projection with a name or title which is more explanatory than merely calling it after the name of the author, or the title of the map, book or atlas in which it was first used.

In this respect the problem of recognising and giving a distinctive label to a map projection is analogous to the way of uniquely identifying the inhabitants of a small Welsh town. In Wales the number of surnames is limited to a handful, like Davies, Evans, Jones, Thomas, and Williams. There are also few christian names. Thus to identify David Jones, the baker, and distinguish him from David Jones, the policeman, and every other David Jones living in the town, it is necessary to introduce a third, descriptive, method of identification ('Jones-the-bread' or 'Dai-book-and-pencil') which give apposite, poetical and frequently scandalous descriptions of the occupation, physical peculiarities or behaviour of each inhabitant. Just as three levels of recognition are needed in Wales, three methods of description and classification are required to identify a map projection. We shall call these:

Aspect *Property* *Class*

The appearance and recognition of map projections

The following projections are illustrated in this book:

Equidistant conical with one standard parallel (Ptolemy), normal aspect	10.07,	208
Equidistant conical with two standard parallels (de l'Isle), normal aspect	10.08,	210
Mercator projection	10.10,	214
Bipolar oblique conformal conical projection	11.03,	231
Fisher's modification of Fawcett's composite equal-area projection	13.06,	279
Kadman's version of the hyperboloid projection	13.08,	285
Polyfocal projection	13.10,	288
Recentred sinusoidal projection	15.01,	314

The reader will find it useful to refer to these in the discussion which follows.

This list indicates some of the methods which are commonly used to identify individual projections. The meaning of some of the words occurring in these titles will become apparent as the reader proceeds. But before we consider the descriptive terminology and classification we must ask the simple question: *How do we recognise a particular map projection?*

Diagnostic features to help recognise a projection

We offer here seven diagnostic features of a projection which ought to be examined. We invite the reader to look at the world graticules in the list above and make notes about the seven features as these affect each map.

(1) Is the world mapped as a continuous feature or are there breaks in the continuity of the map?

Most of the projections in this book represent the whole world on a continuous map, but we find exceptions in Figs 5.01, 13.05, 13.06 and also in Fig. 13.02.

(2) What kind of geometrical figure is formed by the outline of the world or hemispherical map?

The examples include rectangular, circular, elliptical and more complicated outlines.

(3) How are the continents and oceans arranged with respect to the outline and axes of the map?

Many of the projections illustrated provide what we might loosely call a 'conventional' view of the world, which is one to which we are accustomed through frequent exposure to its outlines in atlas maps, books and newspapers. It is the world map in which the equator and the Greenwich Meridian form orthogonal axes and the geographical poles are located at either end of a rectilinear central meridian on the edges of the map. If this conventional arrangement is not apparent can you give any reason why it is not so? Possibly some meridian other than Greenwich has been

used as the central meridian. Possibly the geographic poles are not located on the top and bottom edges of the map.

(4) Are the parallels and meridians rectilinear or curved?

(5) Do the parallels and meridians intersect everywhere at right angles or do oblique graticule intersections occur in some parts of the map?

(6) Are curved parallels or meridians composed of circular or higher-order curves? If the arcs are circular are they also concentric?

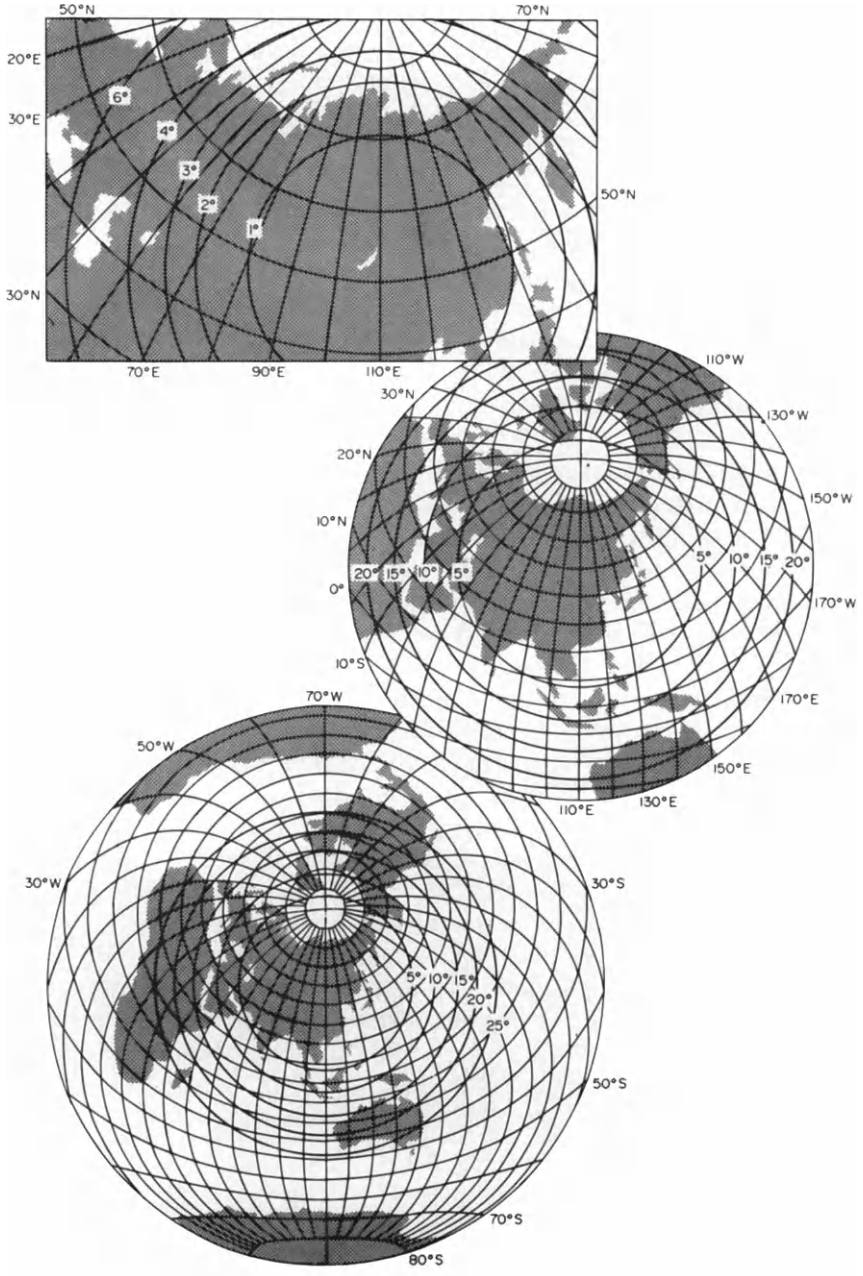
(7) Is the spacing between successive parallels and meridians uniform or variable? If they are not uniformly equidistant does the separation between the parallels increase or decrease from the equator towards the poles? Does the separation between the meridians increase or decrease from the centre of the map towards its edges?

All of these variables can help us to identify a map projection, and most of them will be used in some way or another as the basis of classification. The appearance is of less value in helping us to decide the special property of a projection, for visual inspection often only provides negative evidence. Thus we may state that a map projection with oblique graticule intersections *cannot* be conformal, but this does not mean that all map projections having orthogonal graticules are necessarily conformal. The way in which the parallels are spaced is often helpful in making a more positive guess about special property. Since the area on the earth enclosed between two parallels and two meridians becomes smaller towards the poles, a map projection with small exaggeration in area must also represent this relationship. Comparison of Figs 6.04 and 10.10 indicates that the first of the projections meets this requirement whereas the second does not.

The difficulty of recognition is greatly increased if only part of the world is shown on a map which is arbitrarily bounded by the neat lines. Figure 7.01 illustrates this principle with reference to a map of the

FIG. 7.01 Three different versions of the same aspect of the Azimuthal equidistant projection (No. 11 in Appendix I). The bottom figure shows most of the world represented by means of an oblique aspect of the projection with the origin in latitude 52°N, longitude 110°W. The centre figure shows a hemisphere on the same projection with the same origin. This is the best-known way in which the azimuthal projections are used. The top figure illustrates how only the central portion of an azimuthal projection may also be used to depict a smaller area at a larger scale. In this example it is to be used for an atlas map of the USSR. Note that this kind of map may create difficulties in identification because the characteristic circular outline of an azimuthal projection is truncated by the neat lines of the map. Each of these maps shows isograms for maximum angular deformation (ω). On the two smaller scale maps the isograms are at intervals of 5°, 10°, 15°, 20° and 25°. Greater amounts of angular deformation on the world map are omitted for greater clarity. The larger scale map of the USSR shows isograms for ω at 1° intervals to 5°.

USSR. Clearly the absence of the distinctive circular outlines of the world or hemispheric maps make it more difficult to identify the projection upon which the largest-scale map is based.



The fundamental properties of map projections

A further feature of many of the map projections illustrated in this book is the representation on them of isograms for equal values of maximum angular deformation, ω, or area scale p, or particular scales, μ. This information is not normally shown on maps produced for other purposes, but it provides an alternative method of studying the merits of different projections. Using the methods of interpretation of the distortion characteristics of any map projections, outlined in Chapter 6, pp. 112–118, we may look again at some of the maps to study:

- The nature of the point or line of zero distortion and the location of it with respect to the world or hemispheric outline.
- The location of singular points on the map and how these appear. Usually a singular point is mapped as a line, but sometimes it is removed infinitely far from the origin of the projection so that the map has no real boundary.
- The characteristic patterns formed by the isograms for ω, p, or μ.

We may call these the *fundamental properties* of the projection. Look for similarity of pattern of different map projections (e.g. the comparison of Fig. 6.04, p. 112 with Fig. 10.10, p. 214 shows that both have rectilinear isograms which are parallel to the equator). Look for precisely the same pattern appearing on maps with quite different graticules (e.g. Figs 6.04, 7.02 and 7.03, or Figs 10.02, 10.03 and 10.04).

The first comparison indicates that there are projections with related fundamental properties through different special properties. This suggests that either may serve as the basis for classification. The second comparison indicates that *the fundamental properties of a projection are independent of the graticule.*

We investigate the fundamental properties of three well-known classes of map projection through a description of them in these terms. This introduces us to the three collective names, *azimuthal*, *cylindrical* and *conical*, all of which figured in the titles of the map projections listed on p. 124. In these descriptions we deliberately refrain from referring to the elements of the graticule (equator, poles, parallels and meridians) because we wish to demonstrate that the three fundamental properties are always satisfied by all members of each class of projections, whereas the appearance of the maps may be quite different.

Azimuthal projections

These are sometimes also called 'zenithal projections'. We prefer to use the first name, which has some meaning, and discourage use of the second,

which has none. Some examples of azimuthal projections are illustrated by Figs 1.07, 7.01, 10.02, 10.03 and 10.04.

These projections may be imagined as the transformation to a projection plane which is tangential to the generating globe, as illustrated in Fig. 5.07, p. 90, or intersecting the spherical surface, as in Fig. 5.10, p. 92. We consider the first example here. There is one point of zero distortion, corresponding to the point where the two surfaces meet. In doing this we have by this means reconstructed the definition of a spherical angle illustrated in Fig. 3.03, p. 54, and therefore such projections have the common property that all angles, azimuths in the general case, are correctly represented at the common point. This indicates the reason for the preferred use of the word azimuthal to be the collective name for such maps.

The characteristic outline of the azimuthal map of the hemisphere (and possibly, the whole world too) is circular, and since there is a single point of zero distortion at the centre of the circle, the particular scales increase radially outwards from it in all directions. Consequently the distortion isograms are also circular and concentric from the origin. The singular point of some azimuthal projections is the antipodal point to the origin, which is mapped as the circumference of a circle bounding the whole world map. There are, however, some azimuthal projections which can only be used to map smaller portions of the sphere because the singular point lies at the hemispheric boundary.

Cylindrical projections

Cylindrical projections are illustrated by Figs 6.04, 7.02, 7.03 and 10.10. These projections may be imagined as the transformation to the plane if this is wrapped round the globe in the form of a tangent cylinder, as illustrated in Fig. 5.05, p. 89. Ignoring, for the present, the alternative possibility of the secant cylinder (Fig. 5.08, p. 91) there is a single line of zero distortion corresponding to the great circle of contact, and this is *always represented on the map by a straight line.* Singular points occur at 90° distance from the line of zero distortion on either side of it, and these points are mapped as straight lines which are both parallel to it and of equal length. Consequently the characteristic outline of a world map on a cylindrical projection is rectangular. Distortion isograms are always rectilinear and parallel to the line of zero distortion.

Conical projections

These are also called 'conic projections'. The first of these terms is preferred because the word 'conic' has a different meaning in mathematics (the conic sections) which is totally unrelated to the cartographic usage.

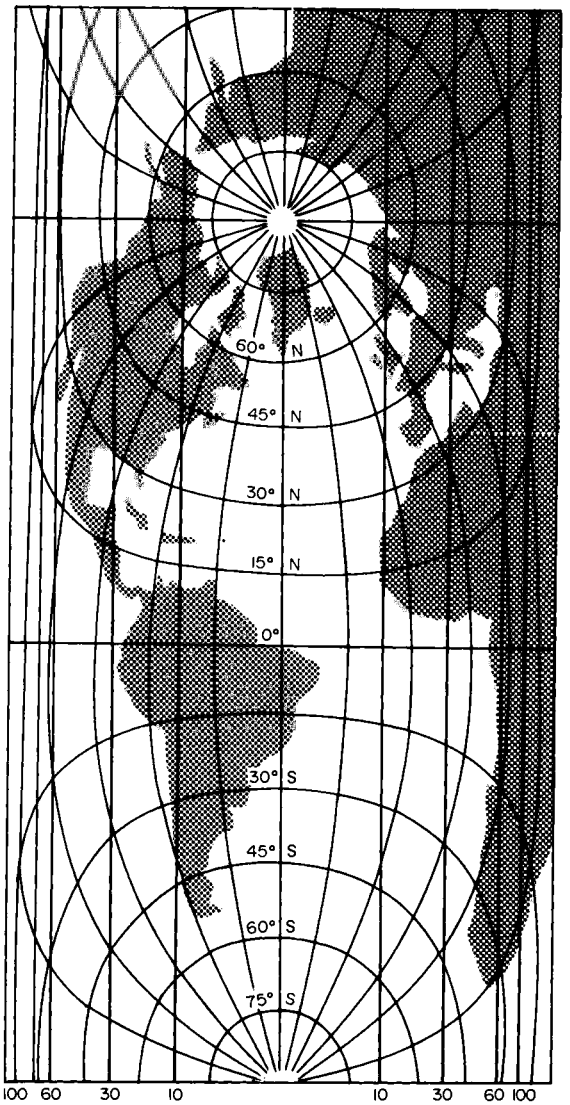


FIG. 7.02 The transverse aspect of the Cylindrical equal-area projection (showing only part of the world), in which the line of zero distortion is the meridian 45°W and its antimeridian 135°E. The map shows isograms for maximum angular deformation (ω) at 10°, 30°, 60° and 100°. These are identical to the corresponding isograms shown in Fig. 6.04, p. 112.

Some examples of conical projections are illustrated by Figs 10.07 and 10.08. This category of projections may be imagined as the transformation from the sphere to the plane through the medium of a cone wrapped round the globe, as illustrated by Fig. 5.06, p. 90, this giving

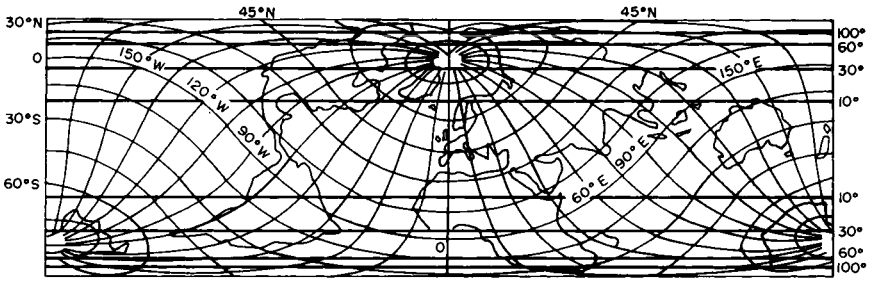


FIG. 7.03 The oblique aspect of the Cylindrical equal-area projection, in which the line of zero distortion is the great circle passing through the points latitude 45°N, longitude 0°, and latitude 45°S, longitude 180°. The map shows isograms for maximum angular deformation (ω) at 10°, 30°, 60° and 120°. Their location is identical to the corresponding isograms in Fig. 6.04, p. 112 and Fig. 7.02, p. 130, these being referred to the rectangular outline of the world map.

rise to a single line of zero distortion corresponding to the small circle of contact, and this is *always represented on the map by a circular arc.* The outline of the hemispherical map is fan-shaped. If the projection is extended far enough to include singular points these are also mapped as circular arcs parallel to the line of zero distortion. The distortion isograms on conical projections are also circular arcs concentric with the line of zero distortion.

The aspect of a map projection

In order to test the validity of these statements the reader should study the three different versions of the Azimuthal equal-area projection illustrated by Figs 10.02, 10.03 and 10.04; also the three different versions of the Cylindrical equal-area projection in Figs 6.04, 7.02 and 7.03. Reference should also be made to Fig. 7.04, pp. 132–133, which illustrates seven different versions of the *Sinusoidal projection*, a member of the *pseudocylindrical class*, as yet undefined.

All three azimuthal projections have the same principal scale and are therefore bounded by circles of equal radius. Figures 6.04 and 7.03 for the Cylindrical equal-area projection are similarly of identical dimensions, but Fig. 7.02 is shorter in length because this map does not show the whole world. Similarly all seven versions of the Sinusoidal projection have identical dimensions, as defined by the lengths of the equator and central meridian in Fig. 7.05(a), which represents the axes of symmetry for the outline of the map in all the examples illustrated.

Thus every version of each projection may be regarded as having an identical outline. Similarly the patterns of distortion isograms are the same for each projection. On the other hand, the appearances of the

parallels and meridians, and therefore the continental outlines, are different on every map.

We use the word *aspect* to indicate the appearance of the graticule. In much English writing on map projections the alternative word in use is 'case'. But the word aspect emphasises the essential ingredients of view and appearance, whereas the word 'case' does not. Moreover, it has many other kinds of unrelated usage in medicine, law, travel and grammar. In order to use a systematic method of defining the different aspects of map projections it is desirable to relate the appearance of the graticule to the fundamental properties of them. We find it convenient to consider a basic threefold subdivision into

- The Normal Aspect;
- The Transverse Aspect;
- The Oblique Aspect.

A cursory glance at Fig. 7.04 indicates that (a) is the simplest pattern of meridians and parallels because all the parallels are straight lines. The

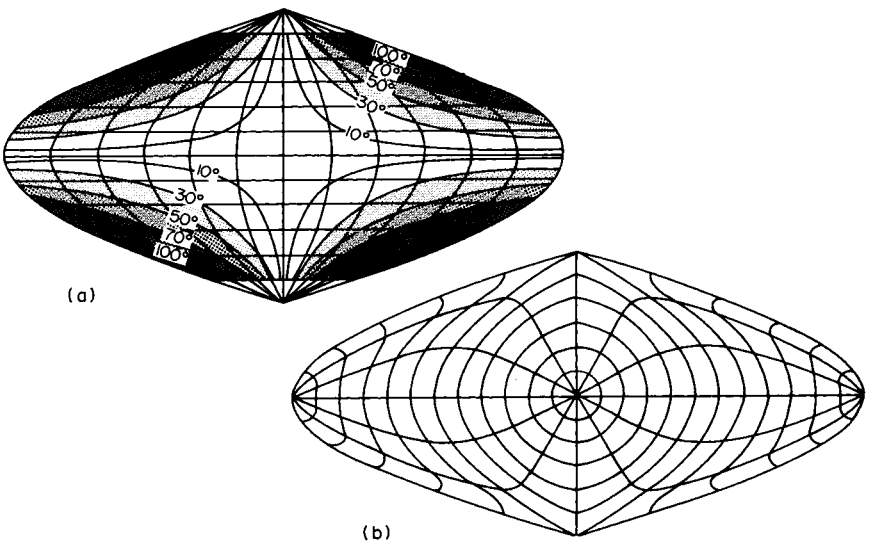
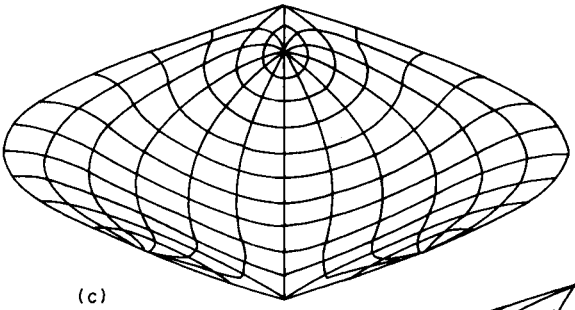
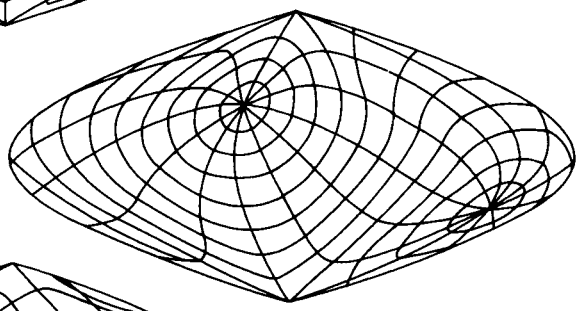


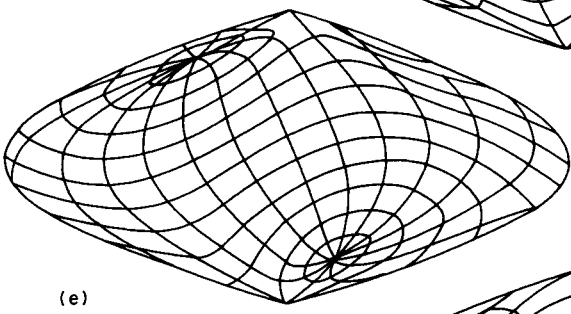
FIG. 7.04 Seven different aspects of the Sinusoidal projection (after Tobler). Figure 7.04(a) is the normal aspect of the projection (No. 30 in Appendix I). This is an equal-area pseudocylindrical projection in which the parallels are equidistantly spaced and the meridians are sine curves. The map shows a 15° graticule and isograms for maximum angular deformation (ω) for 10°, 30°, 50°, 70° and 100°. From the other examples, Fig. 7.05(b) represents the transverse (Wray's first transverse) version. Figures 7.05(c) and (f) represent the simple oblique, in which the minor axis is occupied by the central meridian and is still rectilinear. Figures 7.05(d), (e) and (g) are all versions of Wray's plagal oblique aspect.



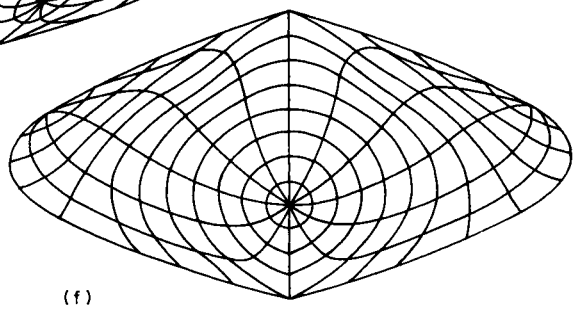
(c)



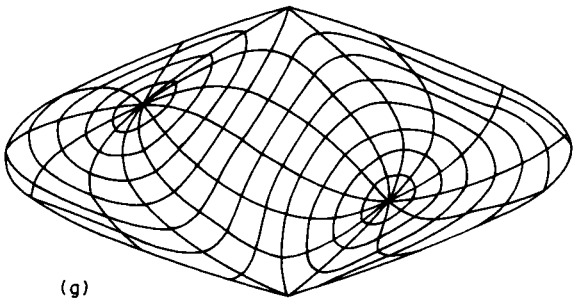
(d)



(e)



(f)



(g)

pattern becomes more complicated in the four examples (d) through (g). Using the convention of the threefold classification, (a) is the normal aspect, (b) is the transverse aspect and all the remainder are oblique aspects of the same projection. Nevertheless there are sufficient distinctive characteristics of the others to suggest that a single category labelled oblique is an inadequate description of them.

The normal aspect

Inspection of each group of illustrations indicates that one of them is geometrically simpler than the others. Thus Fig. 6.04 has a rectilinear network of parallels and meridians, whereas Figs 7.02 and 7.03 both show more complicated patterns of curved parallels and meridians. Moreover in Fig. 6.04 the distortion isograms coincide with certain parallels of latitude, whereas in both Figs 7.02 and 7.03 the isograms intersect the graticule everywhere. In Fig. 10.02 the geographical pole is at the centre of the map – coinciding, therefore, with the point of zero distortion. In this aspect of an azimuthal projection the meridians are rectilinear and the parallels are concentric circles. Moreover, the distortion isograms coincide with certain parallels of latitude. Figures 10.03 and 10.04 indicate more complicated relationships between the isograms and the graticule. In Fig. 7.04(a) the longer axis of the Sinusoidal projection is represented by the equator and the shorter axis by the central meridian. In this particular projection the principal scale is preserved along both of these axes, hence the asymptotic pattern of distortion isograms for ω illustrated in this map. We note that all the parallels are represented by parallel straight lines so that this version is simpler than any of the other diagrams 7.04(b)–7.04(g). We call this the *normal aspect* or *direct aspect* of a projection because there is a direct relationship between the fundamental properties and the graticule, which corresponds to Lee's (1944) dictum that *the direct aspect is always the simplest mathematically*. This rule has also been followed by Wray (1974).

The transverse aspect

We now consider the aspect of the three projections illustrated by Figs 7.02, 7.04(b) and 10.03. In the example of the Cylindrical equal-area projection the central axis of the projection has become the bimeridian formed by a meridian together with its antimeridian, and this is the line of zero distortion. The singular points are the two points on the equator which lie 90° distant from the central meridian, and these are mapped as two equidistant parallel lines of the same length. Thus the fundamental shape of the projection is retained, together with precisely the same pattern of distortion isograms which appeared in the normal aspect. The

graticule is more complicated, but we can see that it is symmetrical about both the central meridian and the equator.

The example of the Azimuthal equal-area projection shown in Fig. 10.03 indicates that the point of zero distortion has been shifted to the equator. This and the central meridian are represented by straight lines which are also two axes of symmetry.

Figure 7.04(b) illustrates the corresponding member of the group of different aspect of the Sinusoidal projection. The longer axis of the projection (what Wray calls the *metaequator*) is now formed by a meridian together with its antimeridian. The equator is formed by two curves, which can be seen, by careful comparison of the two maps Figs 7.05(a) and (b), to correspond to the two meridians 90° from the central meridian of the normal aspect. There are two axes of symmetry which are these two axes of the projection.

These versions may be called the transverse aspect of each projection. The term *equatorial aspect* is also used for this version of an azimuthal projection.

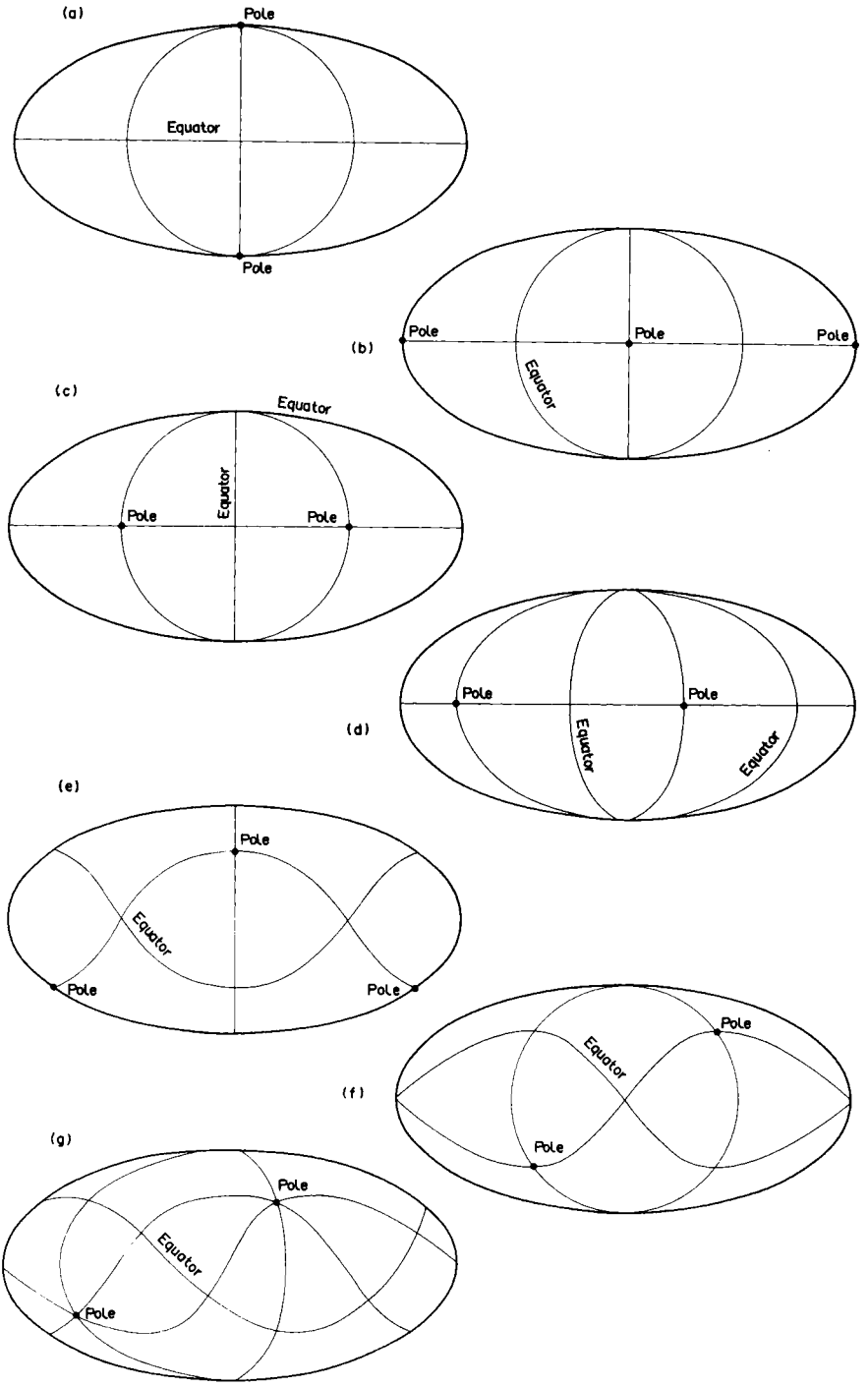
The oblique aspect

The third aspect is shown in Figs 7.03, 10.04 and Fig. 7.04(c)–(g). The large number of different versions illustrated in Fig. 7.04 indicates that there are limitless possibilities of variation. In the Cylindrical equal-area projection the line of zero distortion, which is still the straight line forming the longer axis of the rectangle, corresponds to the great circle passing through the two points in latitude 45°N, longitude 0° and latitude 45°S, longitude 180°. The other axis of the projection is represented in part by the Greenwich Meridian and in part by the antimeridian 180°. There are two singular points, in latitude 45°N, longitude 180° and at 45°S, longitude 0°, which, as before, are mapped as parallel straight lines to form the two longer sides of a rectangle. The pattern of distortion isograms is the same as for the normal and transverse aspects of the projection. The graticule is symmetrical about only one axis, namely the central meridian.

Figure 10.04 illustrates one version of the Azimuthal equal-area projection with the origin in latitude 40°N, longitude 30°W. All the parallels and meridians are curved with the exception of the rectilinear central meridian which also represents the single axis of symmetry. The corresponding examples for the sinusoidal projection are shown by Figs 7.04(c) and (f).

Wray's additional categories of aspect

Shortly after the first edition of this book appeared, the monograph by Wray (1974) was published. He, too, had recognised the complexities of


(a)
Pole
Equator
Pole
(b)
Pole
Pole
Pole
Equator
(c)
Equator
Equator
Pole
Pole
(d)
Pole
Pole
Equator
Equator
(e)
Pole
Equator
Pole
Pole
(f)
Pole
Equator
Pole
(g)
Pole
Equator
Pole

the oblique aspect, and in this work he argued for the acceptance of seven different aspects. In order to describe these adequately he had to introduce many new terms. Thus a transverse projection may be the *first transverse* [Fig. 7.04(b)], *second transverse* or *transverse oblique*, depending upon the position of the geographical poles along the line or curve representing the equator of the projection. A special category of *plagal* or *scalene* oblique aspect projections caters for the skew oblique versions where neither axis of symmetry corresponds to the graticule. Wray's seven aspects of Mollweide's projections are illustrated in Fig. 7.05, and three examples of plagal projections are illustrated in Figs 7.04(d), (e) and (g).

In Wray's terminology these are all examples of the *simple oblique* aspect because each has a rectilinear central meridian, although all other parallels and meridians are curved. The central meridian represents the single axis of symmetry. The only difference between Figs 7.04(c) and (f) is the location of the origin. In (c) $\varphi_0 = 30°\text{N}$; in (f), $\varphi_0 = 60°\text{S}$.

The remaining three examples of oblique aspect Sinusoidal graticules do not have any rectilinear parallels or meridians. Consequently there is no axis of symmetry related to the graticule.

We have described the three aspects of these map projections in detail because we shall find that this is an extremely important concept with considerable practical applications, not only for the design of world or hemispherical maps, as illustrated in the figures cited, but also for maps of individual countries, as indicated by Fig. 7.01 and other examples described in Chapter 12. We repeat the principle that the only difference between a map projection in its different aspects is the pattern of the parallels and meridians, and therefore the location and appearance of the continents and oceans. The fundamental properties of the class of projection and the special properties of the projection itself remain unaltered. Thus we may think of the basic outline of the world map as being a fixed frame of reference, like a picture frame. Behind this frame the picture of the world can be shifted or rotated so that different parts of it occupy the central portion. Since the patterns formed by the distortion isograms are (like cobwebs) related to the frame and not to the picture, these do not change as the patterns of parallels and meridians, continents and oceans are changed. It therefore follows that by careful planning of the aspect of any map we can locate the parts of the earth which have immediate interest in a part of the projection where distortion is small. Conversely the unimportant parts of the world, such as Antarctica on a world population map, may be situated where distortion is greater but

FIG. 7.05 Wray's seven aspects of a map projection applied, in outline, to Mollweide's projection: (a) direct or normal aspect; (b) first transverse aspect; (c) second transverse aspect; (d) transverse oblique aspect; (e) simple oblique aspect; (f) equiskew aspect; (g) plagal or scale aspect. (Source: Wray, 1974.)

does not materially influence interpretation of the map *for the purpose for which it was designed.* We develop these ideas further in Chapters 11 and 12.

In the description of each aspect of the projections studied we have drawn attention to the symmetry of the graticule about one or two axes. This, too, has practical significance when it is necessary to compute the coordinates of graticule intersections. A map having two axes of symmetry is therefore composed of four quadrants, and the coordinates of corresponding graticule intersections differ only from one another by the signs of the (x, y) plane coordinates. This means that such a projection can be constructed from only half the data needed to construct a map which is only symmetrical about one axis. In turn the skew or plagal oblique versions, which have no axes of symmetry related to the graticule, have to be computed in their totality, or four times as much data is required as was needed for the first type of projection.

The classification of map projections

In order to handle the considerable data-base comprising only the map projections which have been described, it is desirable to formulate a system of classification which is, at the same time, collectively exhaustive and mutually exclusive. In other words, the system must include all possible kinds of map projection which have been or are likely to be described. Each projection ought, ideally, to occupy a unique position within the classification system, like every element in the periodic system or each species within the Linnaean classifications of the plant and animal kingdoms. No projections ought to be relegated to categories labelled 'Miscellaneous', 'Conventional' or 'Others', for this creates a kind of garbage can to contain all the varieties of map projection which cannot be conveniently accommodated elsewhere within the system. Reference is often made to projections with 'arbitrary properties' (or 'aphylactic projections') which usually means that these are neither conformal nor equal-area projections. The use of such terms, and the incorporation of such categories within a classification system, is a negative approach with little to commend it.

Only two attempts at classification have really attempted to satisfy these desirable criteria. The first was the so-called 'Linnaean System' described by Maurer (1935), and the second is the *Parametric Classification* of Tobler (1963). Both of these have considerable merit. Maurer's system is the more complicated; Tobler's method of classification has the great merit of being all-embracing and quite simple to understand, but it does not go far enough. The author has therefore taken Tobler's work as the basis for classification, but extended it to produce an ordered hierarchy of *groups*, *classes* and *series*.

The subdivision of all map projections into five groups, A–E, is essentially Tobler's system. This makes use of different combinations of the functional relationships between the map, described in plane rectangular or polar coordinates and the geographic coordinates of the generating globe. Eight such pairs of combinations may be recognised, all of which map the spherical surface continuously. This gives rise to four possible kinds of continuous map. The fifth group represents those *composite* map projections in which there are changes in function and variation in the fundamental properties from place to place. Some examples of these are described in Chapter 13.

In order to simplify understanding of this system of classification we propose that

(1) *Each group, class and series is defined in terms of the normal aspect.* It could be undertaken in more general mathematical terms but it is much easier for the beginner to comprehend the significance of the classification system in respect to the graticule formed by geographical coordinates. Hence we exclude all variations in aspect from the system.

(2) *We define every projection in terms of the simplest, unmodified version.* Thus the modifications introduced by creating two standard parallels or a standard circle do not enter the classification system, nor do the transformations created by introduction of pole-lines or recentred (interrupted) versions of a map which are described in Chapter 13. This may be unrealistic because many of the map projections bearing individual names are modifications of these sorts of other projections. Those who insist that such distinctions are vitally important can easily incorporate yet another classification level subordinate to those used in Table 7.02, p. 148. However, the object of the classification system demonstrated there is that it should be relatively simple. This does not mean that detailed information concerning aspect and modification should be omitted from the description of a projection on a map. We shall see the importance of this in Chapter 19, when we consider the methods of transformation which may be used after digitising source maps. In order to make the initial subdivision of all map projections into the four groups, A–D, we make use of the functional relationships between plane and geographical coordinates which were introduced early in Chapter 5. An understanding of this notation, as given in equations (5.01)–(5.04), pp. 80–81, is essential. Therefore the reader who skipped that part of Chapter 5 should now refer to it. In order to explain the system in terms of the appearance of the graticule in each of the four groups we must also be explicit about the origins of the plane (x, y) and (r, θ) coordinate systems and also define the orientation of the axes or initial line with respect to parallels and meridians.

Plane representation by cartesian coordinates

We specify that, for a map projection to be defined by plane rectangular coordinates, the origin of the system is located on the equator at its intersection with a selected central meridian. This may be the Greenwich Meridian, as shown in some of the illustrations, but this is not an essential condition. The abscissa of the plane coordinate system coincides with the equator and the ordinate with the central meridian. It therefore follows that x varies mainly with longitude whereas y varies mainly with latitude. In equations (5.01) and (5.02) both x and y vary with both latitude and longitude. This is the general case which we list below as A. However, it may be simplified in three different ways, where either x or y or both x and y vary with longitude or latitude only. This gives rise to the following expression:

$$\left.\begin{aligned} x &= f_1(\varphi, \lambda) \\ y &= f_2(\varphi, \lambda) \end{aligned}\right\}(A) \qquad \left.\begin{aligned} x &= f_1(\lambda) \\ y &= f_2(\varphi, \lambda) \end{aligned}\right\}(B)$$

$$\left.\begin{aligned} x &= f_1(\varphi, \lambda) \\ y &= f_2(\varphi) \end{aligned}\right\}(C) \qquad \left.\begin{aligned} x &= f_2(\lambda) \\ y &= f_2(\varphi) \end{aligned}\right\}(D)$$

The graphical appearance for these functions is illustrated in Fig. 7.06. Clearly (A) represents the general case expressed by (5.01) and (5.02) and (D) is the simplest where x and y are functions of only one variable, namely longitude and latitude respectively. Where x or y, or both coordinates, vary with both latitude and longitude, each parallel of meridian must be represented by either an inclined straight line or a curve. The only exception to this rule are the axes of the plane coordinates where $f_1(\varphi, \lambda) = 0$ or $f_2(\varphi, \lambda) = 0$ and both the equator and central meridians are represented by perpendicular straight lines. Hence a function of the sort

$$x = f(\varphi, \lambda)$$

indicates a curve, the exact nature and location of which is, as yet, unspecified.

A parallel of latitude, by definition, represents the circumference of a small circle on the globe along which φ is constant. Similarly a meridian represents the great semicircle along which λ is constant. If we specify that $x = f(\lambda)$ or that $y = f(\varphi)$, this means that any line depicting a constant value of φ or λ can only have one value for x or y. In other words, if $x = f(\lambda)$ each meridian will be represented by a straight line which is

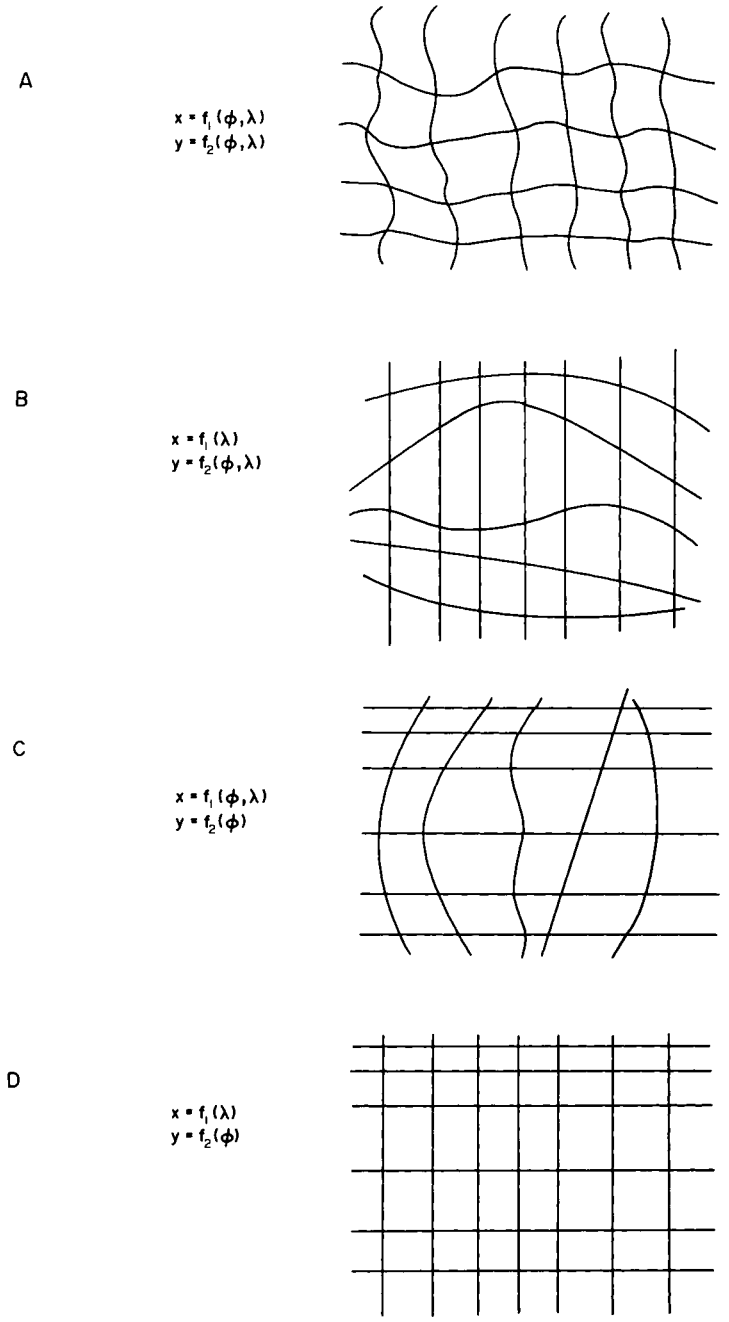


FIG. 7.06 Diagrammatic representation of the geometrical meaning of the four possible function relationships between geographical coordinates and plane cartesian coordinates. (Source: Tobler, 1963.)

parallel to the central meridian. Similarly if $y = f(\varphi)$ each parallel will be represented by a straight line which is parallel to the equator. We therefore have four basic types of map projection which may be defined by plane cartesian coordinates.

Group A comprises the general case where both the parallels and meridians are composed of curves, as illustrated in Figs 1.05, 5.02 and 6.08. This group is known to most writers as the *polyconic* class of projections, although we must comment that this is an unfortunate choice of name because it is also applied to only one projection.

Group B contains projections which have rectilinear meridians which are parallel to the central meridian, and curved parallels. This group has few named members and contains few projections which have any practical use in conventional cartography. The group does include certain projections which have other kinds of use, for example as graphic aids to the solution of spherical triangles in astronomical navigation.

Group C contains projections which have curved meridians and parallels composed of parallel straight lines. These are called *pseudo-cylindrical* projections illustrated, for example by Figs 6.07, 7.04, 7.05, 13.05 and 15.01.

Group D is the simplest of the four categories and must comprise projections which comprise two families of parallel straight lines. These are the *cylindrical* projections already introduced in Fig. 6.04.

Plane representation by polar coordinates

We employ similar arguments to subdivide the possible varieties of map projections which are more conveniently described in terms of plane polar coordinates. We specify that the origin of the system is located at or near one of the geographical poles, and that the initial line coincides with the central meridian. Thus the radius vector, r, represents the distance from the origin to a parallel of latitude and is therefore a measure of colatitude. However, this is a function of latitude so we may retain the convention that $r = f(\varphi)$. The vectorial angle, θ, is related to the spherical angle measured at the geographical pole; therefore θ is predominantly a measure of longitude. However, we have created some uncertainty in this specification by stating that the origin of the coordinate system is located 'at or near' one of the poles. This creates further complication which means that for each of the four possible pairs of functions there exist two possibilities. The first is where the origin of the polar coordinates is actually at the geographical pole; the second is where it is located at some *vertex*, which is a point on the prolongation of the polar axis beyond the spherical surface. Bearing in mind that we have this dual interpretation, the four pairs of functions may be written in the form:

$$\left.\begin{aligned} r &= f_1(\varphi, \lambda) \\ \theta &= f_2(\lambda) \end{aligned}\right\}(B)$$

$$\left.\begin{aligned} r &= f_1(\varphi, \lambda) \\ \theta &= f_2(\varphi, \lambda) \end{aligned}\right\}(A) \qquad \left.\begin{aligned} r &= f_1(\varphi) \\ \theta &= f_2(\lambda) \end{aligned}\right\}(D)$$

$$\left.\begin{aligned} r &= F_1(\varphi) \\ \theta &= f_2(\varphi, \lambda) \end{aligned}\right\}(C)$$

As before, $f(\varphi, \lambda)$ indicates a curved parallel or meridian. Where r is a function of latitude only the parallels are represented by concentric circular arcs. Where θ is a function of longitude only, the meridians are straight lines converging towards the origin of the coordinate system. The functions represented by (A) correspond to projections in which both the parallels and meridians are curved, and may be grouped with the cartesian group A as polyconic projections. The intermediate functions of group (B) have rectilinear meridians and curved parallels, which may be grouped with the cartesian group (B) also. The two remaining groups (C) and (D) both have $r = f(\varphi)$ and therefore have parallels represented by concentric circular arcs. If the origin is located at the geographical pole, the parallels are represented by the circumferences of circles which have their common centre at this point. In group (D) the meridians radiate as straight lines from the origin, defining the *azimuthal* class of projections. In group (C) the meridians are curved and are called *pseudoazimuthal* projections. This is another unfamiliar class. See Arden-Close (1952) and Snyder and Voxland (1989) for illustrations of one of the few examples of this class which has been described. On the other hand, if the origin of the polar coordinates is situated at some vertex, the parallels are again represented by concentric circular arcs but cannot form a complete circumference. The resulting shape of the projection depends upon the shape of the meridians. In group (D) the meridians are rectilinear, giving rise to the characteristically fan-shaped *conical* projections. In Group C the meridians are curved, producing the bell-shaped *pseudoconical* projections.

We have now created four groups (A)–(D) with combinations of functions which include all possible ways in which the spherical surface can be mapped *continuously* upon a plane. Within these groups there are seven named classes, three of each in groups (C) and (D), together with the word polyconic, which is applied to the entire group (A). This stage of classification may be illustrated diagrammatically, as in Fig. 7.08.

Separation of the parallels

Thus far we have not specified any particular condition concerning the spacing of the parallels; we have only stated that y is some function of latitude. However, if the reader has studied the projections illustrated, as

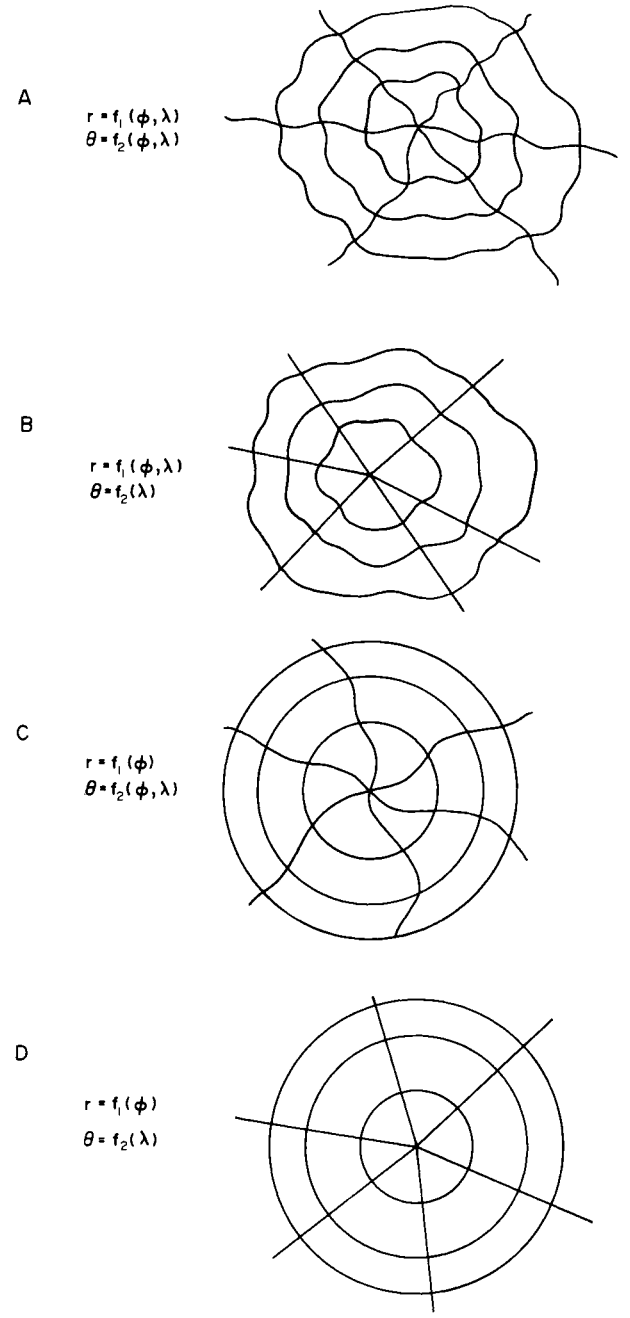


FIG. 7.07 Diagrammatic representation of the geometrical meaning of the four possible functional relationships between geographical coordinates and plane polar coodinates. (Source: Tobler, 1963.)

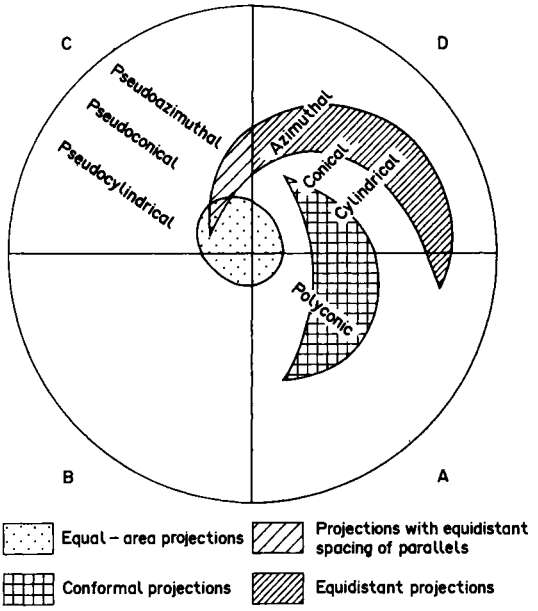


FIG. 7.08 Venn diagram illustrating the relationship between Tobler's four groups (A)–(D), the subdivision of the groups into named classes, and indicating the relationship of certain special properties of map projections to the system of classification. Study of this diagram indicates, for example, that conformal map projections are confined entirely to groups (A) and (D) and that equidistant map projections are confined to group (D). The study will find it instructive to plot the locations of the projections listed in Appendix I upon an enlarged copy of this diagram.

suggested at the outset of this chapter, it will be appreciated that the spacing can vary in three different ways:

- the separation between the parallels *decreases* with increasing latitude from the equator towards the poles;
- the separation between the parallels *remains constant* for all equal increments of latitude;
- the separation between the parallels *increases* with increasing latitude from the equator towards the poles.

The first and third conditions can vary in an infinitely large number of ways, but the second cannot change.

For the cylindrical and pseudocylindrical classes these variations may be conveniently expressed in terms of trigonometric functions of latitude. For example, the sine of an angle varies in such a way that the difference between the sines of two angles close to 0° is greater than the sines of two corresponding angles near 90°. This can be seen from the numerical

variations to be found in the interpolation columns of a table of natural sines in any set of trigonometric tables. Thus for the first case of decreasing separation we may write

$$y = f(\sin \varphi) \tag{7.03}$$

and we will study the specific example of the Cylindrical equal-area projection (Fig. 6.04) in which the parallels crowd together in high latitudes. For this projection we may write

$$y = R \,.\, \sin \varphi \tag{7.04}$$

and determine numerical values for the ordinate for the condition that $R = 1$. Since we have put $R = 1$, the values for y represent a table of natural sines of the angle φ. The column headed δy lists the differences between successive values of y. This column shows that, for a difference in latitude of 15°, the distance between the parallels $\varphi = 0°$ and $\varphi = 15°$, is 0·2588 units, but the distance between $\varphi = 75°$ and $\varphi = 90°$ is only 0·0341 units.

The converse case is

$$y = f(\tan \varphi) \tag{7.05}$$

which indicates that the spacing between the parallels increases from the equator towards the poles. Using a set of natural tangents the reader is invited to construct the table corresponding to Table 7.01, which demonstrates this increase. It is the ordinate for the *Central perspective*

TABLE 7.01 *Values for the ordinate and the difference between successive values for y; Cylindrical equal-area projection (Lambert)*

Latitude	y	δy
0°	0·0000	
		0·2588
15°	0·2588	
		0·2412
30°	0·5000	
		0·2071
45°	0·7071	
		0·1289
60°	0·8660	
		0·0999
75°	0·9659	
		0·0341
90°	1·0000	

cylindrical projection which is a curiosity of little practical value. Since $\tan 90° = \infty$ it follows that the geographical poles cannot be shown on a map because they lie infinitely far from the equator on this map.

The intermediate case is where the parallels are equidistantly spaced. Then we may write

$$y = R \,.\, c \,.\, \varphi \tag{7.06}$$

where φ is measured in radians and c is a constant.

Where $c = 1$ the separation of the parallels corresponds to the arc distance between them on a globe, because this is just another way of expressing equation (3.10). Moreover, in certain classes of projection in group (D) this also corresponds to making the particular scale $h = 1$ and creating a map having the special property of equidistance. However, we must note that equal separation of the parallels does not necessarily ensure that the projection is equidistant. Pseudocylindrical projections frequently have equidistantly spaced parallels [e.g. Fig. 7.05(a)] but $h = 1$ along the central meridian only. All the other meridians are curved and therefore h varies from point to point on the map.

We employ the three principles relating to the spacing of the parallels in the classification. However, the *sine series* and *tangent series* cannot be applied as descriptive terms for all classes of projections. Therefore we use *decreasing separation*, *increasing separation* and *equidistant parallels*, as being an all-embracing form of subdivision. In each case we mean the change in the separation of the parallels proceeding from the equator towards the poles.

A recent paper by Nyerges and Jankowski (1989) represents a comparative study of the methods of classification which have been used in recent work. Essentially this is a comparison of Goussinsky's method, used by Richardus and Adler (1972); the method described here which is unchanged from the first edition of this book and the method used by Snyder (1987a) to classify the projections employed by the US Geological Survey. Nyerges and Jankowski failed to recognise that there are several gaps in the system given in Table 7.02, which have been left because there are no particularly useful projections to be listed therein. For example, the pseudoazimuthal projections are regarded by them as only having one member; which is Wiechel's projection, an equal-area member of the category in which the spacing of the parallels increases towards the poles. There is no *a priori* reason why there should be no members of the other categories in which the parallels are equidistant or the separation decreases. It is just that nobody has found a use for them.

Table 7.02 indicates the proposed system of classification including the projections which have been illustrated and those which are given in Appendix I.

TABLE 7.02 *Classified list of the principal map projections. This classification is based upon: (1) Tobler's groups, A, C, and D, (2) named classes, (3) series formed by spacing of parallels. It refers to the normal aspect in each group and class and does not include any modified projections*

Group	Class	Series: Decreasing separation	Series: Equidistant spacing of parallels	Series: Increasing separation
D	*Cylindrical* $x = f_1(\lambda)$ $y = f_2(\varphi)$	Cylindrical equal-area (Lambert) projection (E)	Plate Carrée (Q)	Mercator's (C) (L) Braun's perspective cylindrical
D	*Azimuthal* $r = f_1(\varphi)$ $\theta = f_2(\lambda)$	Stereographic (C) Gnomonic (O) Minimum-error (Airy) Projection (M) Breusing's (Geometrical Mean) azimuthal Breusing's (Harmonic Mean) azimuthal (M)	Azimuthal equidistant (Postel) (Q)	Azimuthal equal-area (Lambert) (E) Orthographic
D	*Conical* $r = f_1(\varphi)$ $\theta = f_2(\lambda)$	Conical equal-area with truncated pole (E) Conical equal-area with point pole (Lambert) (E)	Equidistant conical (Ptolemy) (Q) Minimum-error conical (Murdoch III) (M)	Conformal conical (Lambert) (C)
C	*Pseudocylindrical* $x = f_1(\varphi, \lambda)$ $y = f_2(\varphi)$	Mollweide's (E) Pseudocylinidrical equal-area with elliptical meridians (Fournier II) (E) Parabolic (E)	Sinusoidal (E) Pseudocylindrical with elliptical meridians (Apianus II) Polyhedric	
C	*Pseudoazimuthal* $r = f_1(\varphi)$ $\theta = f_2(\varphi, \lambda)$			Equal-area pseudoazimuthal (Wiechel) (E)
C	*Pseudoconical*		Bonne's (E)	
A	*Polyconic* $x = f_1(\varphi, \lambda)$ $y = f_2(\varphi, \lambda)$	Hammer–Aitoff (E)	Simple polyconic Aitoff's Tripel (Winkel)	

Key to special property: E = equal-area; Q = equidistant; C = conformal; M = minimum-error (of that class); L = loxodromic (rectilinear rhumb lines); O = orthodromic (rectilinear great circles).

Naming of map projections

A variety of different map projections have been mentioned in this chapter. Some of them are named after the supposed inventor or originator of the projection, such as *Mollweide's projection*, *Aitoff–Wagner projection*, *Bonne's projection*, *Mercator's projection*. Less commonly projections are named after the book or atlas in which they were first used. The *Oxford projection* and *The Times projection* are examples of this usage. A third group are named according to specific mathematical features of the graticule. The *Sinusoidal projection* is so named because the meridians of the normal aspect are sine curves. Many projections have alternative names (*Sanson-Flamsteed's projection* = *Sinusoidal projection*) and many have no proper names.

In giving a map projection a suitable name, the following rules have evolved:

(1) Certain names should be inviolate because of their long history of international use. These include the names of the azimuthal projections originally described in antiquity by Greek and other geometers, such as *Stereographic*, *Gnomonic* and *Orthographic*. It further includes some personal names used for extremely well-known projections with a long history and considerable practical importance, for example, *Mercator's projection*, *Bonne's projection* and *Cassini's projection*. This system of nomenclature becomes unworkable after the late eighteenth century with the prolific inventiveness of J. H. Lambert, who described half a dozen projections all of which are still important in practical cartography. Any one of these might justifiably be called 'Lambert's projection', but each needs additional description in the title to facilitate recognition. Thus we see the emergence of a second method of descriptive name:

(2) The great majority of projections ought to be referred to in terms of: (a) aspect (if other than normal); (b) class; (c) special property; (d) name of originator; (e) nature of any modifications. Thus we may distinguish between the *Cylindrical equal-area* (*Lambert*) and the *Cylindrical equal-area (Behrmann) with standard parallels in 30°N* and *S*. Although this is a bit of a mouthful, it is a precise description of the projection, its origin and modification, all of which information may be important in using this as a source map for further compilation and for measurement purposes by a user of the map. Snyder (1987c) has examined the names of projections which have been used in several modern atlases, and has commented adversely upon the lack of precise information which is available in many of these examples.

Notwithstanding the obvious need for names which uniquely identify individual projections, there are many alternative names in use which complicate the process and are particularly frustrating to the beginner

who has enough unfamiliar terms to learn without any unnecessary duplication of them. Despite the effort put into standardisation of terms during the 1960s, new versions of old concepts still appear, as exemplified by the two new names which have appeared in the Open University television programme and Royal Geographical Society newsletter already mentioned. Of these, 'Lambert Horizontal' is the normal aspect Cylindrical equal-area projection so that use of the name Lambert is correct. However, the word 'horizontal' has no special meaning in cartography, nor, for that matter, in mathematics, which could be construed as helping the user to identify this particular map projection and distinguish it from all others. The name 'great circle map' is used for the Azimuthal equidistant projection. In this respect it might be argued that because the special property of equidistance combined with the fundamental properties of the azimuthal projections allows the user to measure great circle arc distances *from the centre of the map*, and because these particular great circles are rectilinear, such measurements contain no errors which are attributable to the projection itself. However, the specific mention of great circles with respect to map projections normally refers to the *orthodromic* special property; that *all* great circle arcs are represented on the map by straight lines. This is rigorously satisfied by the *Gnomonic* projection, which is also an azimuthal projection, although several other projections approximate to the orthodromic special properties. In fact the term *Azimuthal equidistant projection* serves adequately to describe the so-called 'great circle projection', and *Gnomonic projection* has been a good enough name for the other since about 500 BC. In the example of the Royal Geographical Society logo, first used in the 1920s and finally replaced in 1989, the projection is an oblique aspect azimuthal projection with origin on the Greenwich Meridian in latitude 30° North, which may well be the perspective azimuthal projection attributed to Sir Henry James and named after him.

On the use of personal names

A major problem in nomenclature is the extent to which personal names should intrude into the scheme. We have seen that some personal names have a long and respectable history, allowing identification of most of the projections devised between the fourteenth and eighteenth centuries. After that time, however, there are so many names, and sometimes so many projections, to be ascribed to one person that the memory is strained by the profusion of them. We have become accustomed to refer to the six pseudocylindrical projections described by Max Eckert in 1906 in the form Eckert I through Eckert VI. The same rule works well enough for names line Winkel (three projections), van der Grinten (four projections) or even Schjerning (six projections). However the total of six seems to be

as many as memory can conveniently hold. Consequently we have difficulty in remembering the characteristics of the dozen or so associated with Ginzburg. He used a different method, distinguishing most of them by the name of the institute where he worked and the year each graticule was first used. For example, Ginzburg called each of the world polyconic graticules devised by him a *Polikonicheskaya proektsiya TsNIIGAiK* followed by a date. The initials TsNIIGAiK stand for the Central Scientific Research Institute in Geodesy, Air-survey and Cartography in Moscow. Thus the Russian literature and atlases distinguish *Polikonicheskaya proektsiya TsNIIGAiK (1939–49)* from that of 1950 and 1954 as the three projections used for world maps which were called Ginzburg (IV), Ginzburg (V) and Ginzburg (VII) respectively in Maling (1960).

Notwithstanding the obvious objections to using personal names, they are likely to persist simply because of the greater ease of association of an abstract concept with an easily identified name. The easier it is to remember, or the more alliterative the name, the more likely is this form of identification to stick. A notorious modern example is that of *Peters' projection*, which is a version of the Cylindrical equal-area projection which was appropriated by Arno Peters in 1973 for use with a world map. This name has stuck, evidently irremovably, to this version of the Lambert graticule notwithstanding the fact that Peters' name only became associated with it more than a century after Gall first described it. However the title *Gall's Orthographic projection* slips less easily from the tongue and is more difficult to remember. Sometimes, however, it is almost impossible to forget a name, once heard; *Boggs' Eumorphic* is the classic example of this.

CHAPTER 8

Practical construction of map projections

'Why,' said the Dodo, 'the best way to explain it is to do it.'
Lewis Carroll, *Alice in Wonderland*

Introduction

The compilation of every map should commence with plotting some kind of grid or graticule, for this is the geometrical framework upon which it is based, and which determines the quantitative or positional accuracy of everything shown upon it. At the level of map use the network of lines forms an important frame of reference which can be used to define position and serve as a form of control over both linear and area measurements by permitting evaluation of deformation of figures having known dimensions and areas. For example, the amount of distortion of the paper upon which a map is printed can be determined by measuring known distances between the grid or graticule intersections. The various forms of geometrical control which may be employed in quantitative map use have been described by Maling (1989).

Representation of grids and graticules on maps

At the larger map scales, greater than 1/10 000, the framework used on the map is almost invariably a grid. Nowadays most large-scale maps have neat lines which are grid lines and the sheet numbering system is also based upon the grid. At scales smaller than 1/1 000 000 only the graticule is shown and often the neat lines are a pair of parallels and meridians enclosing a spherical quadrilateral or *quadrangle*. At scales intermediate between these extremes, corresponding to most topographical map series, both the grid and graticule are probably shown. Where the grid and graticule both appear on the map, the draughting specification usually calls for the representation of certain grid lines in full. The graticule is then only depicted by small crosses at the points of intersection of selected parallels and meridians and by subdivisions of the

TABLE 8.01 *Spacing of grid and graticule commonly found on published maps at different scales. Data for large-scale and topographical scales indicate Ordnance Survey practice.*

Scale	Grid separation (km)	Graticule separation (degrees and minutes)
1/2500 and larger	0·1	—
1/10 000	1·0	1′ (margin only)
1/25 000	1·0	*
1/50 000–1/63 360	1·0	1′ (margin only) 5′ (optional crosses)
1/250 000	10·0	1′ (margin only) 30′ (optional crosses)
1/625 000	10·0	10′ (margin only)
1/1 000 000	–	1° 1′ (margin only)
1/2 000 000	–	1°
1/5 000 000	–	2°
1/10 000 000	–	5°
1/20 000 000	–	5–10°
1/50 000 000	–	10–15°
1/100 000 000	–	10–20°
Smaller than 1/100 000 000	–	15–20°

* The only graticule information on the 1/25 000 OS series is a statement of the geographical coordinates of the sheet corners.

margin round the neat lines of the map. Often these crosses are omitted from parts of the map where they coincide with, and might confuse, other detail.

The spacing of the grid lines and graticule depends upon the scale and purpose of the map. Table 8.01 shows the kind of intervals which are commonly found on land maps and in atlases. Nautical and other navigation charts, which are used for precise plotting of position and direction, frequently have closer network of parallel and meridians than are found on maps of equivalent scale.

On a multicoloured map the graticule is generally printed in the colour of the base plate (marginal information, settlement and boundaries) which is either black or dark brown. The graticule is nearly always represented by lines of gauge 0·1–0·2 mm (4–6, measured in units of one-thousandth of one inch). On national topographical maps the grid is also often represented in the colour of the base plate. Grid lines are usually continuous. The *rouletted grid* in which the lines are composed of small closely spaced dots is now almost wholly obsolete. The draughting specification may also require emphasis of certain integer grid lines (usually at every 10 km or 100 km) by a wider gauge than the remainder. On military maps the grid is frequently printed in some colour other than black. The use of a different colour facilitates rapid location of grid lines with respect to

other map detail. Moreover, the use of a distinctive colour for each grid zone provides a means of distinguishing between two overlapping grids where these have to be shown on the same map sheet.

Since the grid or graticule represents the mathematical framework upon which the whole of the rest of the map is based, it follows that grid or graticule intersections should be plotted with great accuracy, and the component lines ought to be fair drawn with considerable care. Most of the map accuracy specifications which have been drawn up for the purposes of legal contract or guarantee refer to *planimetric error* as the displacements of map detail compared with their surveyed positions, both measured relative to the grid or graticule. For further details see Maling (1989). The graphical work of navigation is done with reference to the parallels and meridians on the printed chart. It follows, therefore, that the grid or graticule of a map ought to be plotted 'without sensible error', corresponding to standard error in position of $\pm 0{\cdot}1$ to $\pm 0{\cdot}15$ mm. This is an exceedingly high standard to achieve, and it is therefore necessary to examine the practical ways in which it may be accomplished.

Reference to the illustrations of map projections which appear in this book indicates that certain lines are straight; others are arcs of circles, ellipses and other conic sections. Some projections contain higher-order curves and these may have reverse curvature or even cusps, where curvature is discontinuous at a point. Only a few instruments are available for drawing these lines. Obviously a straight-edge assists drawing straight lines and a half-set or beam compass can be used to construct circular arcs, but this is practically the limit to the instruments which can be used to draw curves to a particular specification. More complicated kinds of mechanical *trammel* have been designed and used to draw the other conic sections (ellipses, parabolae and hyperbolae) but these instruments are quite rare and they are not particularly reliable. Thus, in the absence of a suitable computer/plotter combination which will allow automatic plotting and drawing of the curves, it is necessary to use some kind of template to be laid through the points representing graticule intersections and guide the ruling pen or scribing tool. These may be flexible, *splines* or *flexicurves*, or they may be rigid drawing aids known as *French curves*, *ship curves* or *Copenhagen curves*. We will use the word 'curves' to mean any of these rigid varieties. Splines and curves have to be fitted by trial and error until part of the ruling edge passes through a succession of plotted points. Consequently a graticule composed of curves can be drawn only after the individual graticule intersections have been located on the plotting sheet. Hence the first stage in constructing any map projection is to plot the positions of the graticule intersections at the required scale. The second stage is to draw the individual parallels and meridians through these points. Finally, of course, the topographical detail of coastlines, rivers, roads, railways and towns have to be transferred from the source

maps to the new plotting sheet, fitting it within the control now offered by grid squares or graticule quadrangles.

Location of graticule intersections

A variety of graphic and semi-graphic techniques have been used for construction. It is necessary to emphasise that there is a considerable difference between the wholly geometrical methods of construction described in most of the elementary textbooks on map projections published in the English language and the techniques which are used by professional cartographers.

Geometrical construction

For some projections the methods of plotting can be carried out entirely using ruler and compasses; indeed the whole geometrical construction can be accomplished without having to make any calculations apart from the initial determination of the scale of the intended map. The construction of each projection is unique and therefore it must be learnt in advance. The elementary textbooks are full of such recipes, and many students of the subject are led to imagine that the study of map projections comprises learning these by rote. The author believes that this approach to the subject is wrong and it is a waste of time, for the following reasons:

- Geometrical construction tends to be progressive so that the work proceeds 'from the part to the whole' without much opportunity for checking the accuracy of the construction and is almost always concerned with graphical *enlargement*. This means that inevitable small errors of plotting introduced at the initial stages of construction accumulate to quite large errors in the final positions of the meridians and parallels. Elsewhere in this chapter (pp. 166–171) we examine similar problems in the discussion of suitable methods for constructing a master grid when this has to be done graphically.
- Geometrical construction has to be limited to the preparation of very small-scale maps. This is because there are few straight edges, beam compasses and splines of length greater than 1 metre, and even quite small-scale maps need construction arcs which are of greater radius than this. Instruments such as beam compasses are quite impracticable, and lack precision when used to describe circular arcs of radii in excess of 2 metres. Also some construction points may have to be located beyond the boundary of the intended map, often a long way from the centre of the projection. This requires a very large plotting table and can be extremely wasteful in the consumption of draughting film. For example the transverse aspect of the

Stereographic projection, illustrated in Fig. 1.07, p. 15, can be constructed entirely from straight lines and circular arcs using ruler and compasses. However, the radii of the meridians close to the central meridian are so great that, even at a very small scale, it is impracticable to attempt to locate their centre and describe these arcs by beam compass. There isn't space on the table, the beam compass is too short and there usually isn't sufficient plastic left in the roll to draw such large radius arcs.

- The final and most important objection of all is that most of the really useful map projections cannot, in any case, be constructed by geometrical methods.

Construction by coordinates

The method to be described comprises the calculation of the plane cartesian coordinates of each graticule intersection, plotting these upon a *master grid* and finally joining the plotted points by smooth curves to represent the parallels and meridians. The master grid may be preprinted, duplicated or constructed especially for the map. The coordinates may be read from tables which have already been published. If there are no suitable tables, numerical values have to be determined using the projection equations, for example those listed in Appendix I, pp. 430–441. Figure 8.01 indicates the various ways in which the grid or graticule may be plotted using traditional methods.

Construction of a map projection by plotting coordinates

Let us suppose that we wish to construct a fairly complicated map projection, for example *Briesemeister's projection*, for the whole world at a scale of 1/40 000 000 showing the parallels and meridians at intervals of 15°. A smaller-scale version of this map with a 20° graticule is illustrated in Fig. 8.02 and the coordinates required for construction are given in Appendix II on p. 442.

This projection was first described by Briesemeister (1953, 1959) and has been used as an equal-area base for world maps in many American Geographical Society publications such as their *Atlas of Diseases*, and in many United Nations reports. It is one of many possible modifications to the Hammer–Aitoff projection, from which it differs by having the ratio 1:1·75 between the axes instead of 1:2 in the parent projection. It is also only used in the simple oblique aspect with origin in latitude 45°N, longitude 15°E. See Maling (1962) for an account of its mathematical derivation.

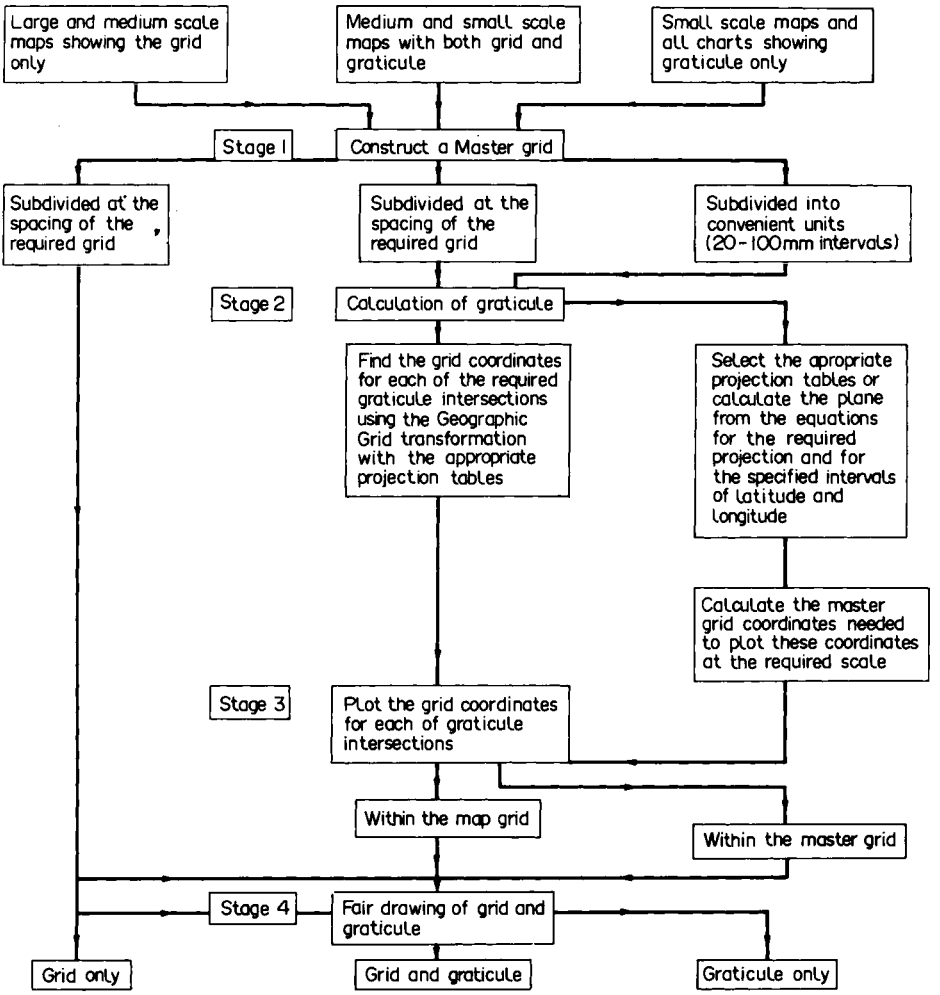


FIG. 8.01 Flow diagram illustrating the methods of constructing the geometrical framework for a map using the conventional draughting methods and without the direct aid of computer graphics. (Source: Maling, in ICA, 1984.)

The required equipment and data are:

- A master grid showing a 5 mm net on a sheet of dimensionally stable plastic of suitable dimensions.
- Tabulated values of the rectangular coordinates for the 15° graticule for this projection.
- A pocket calculator to convert the tabulated coordinates into the *master grid coordinates* required to plot the map at the required scale.

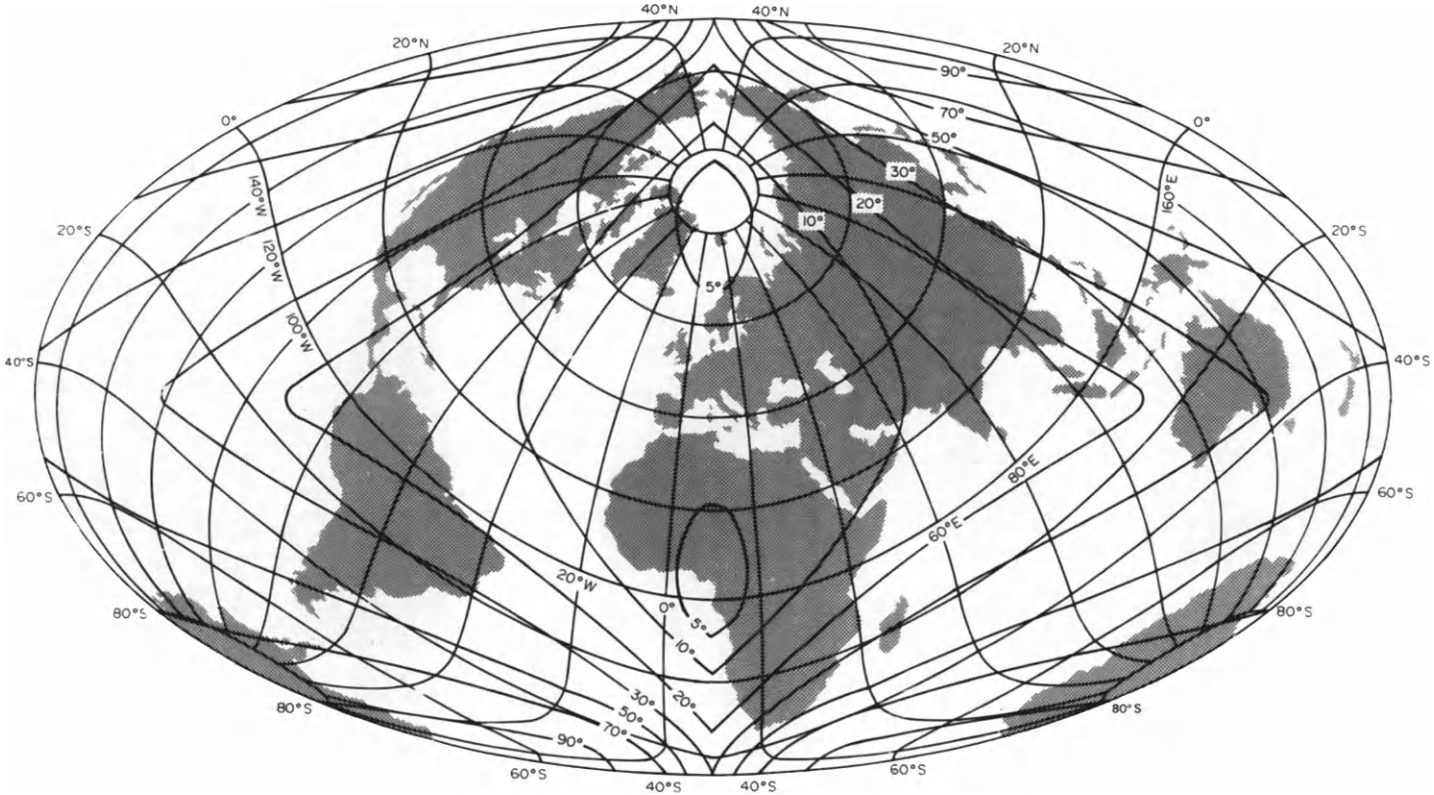


FIG. 8.02 Map of the world in Briesemeister's projection (No. 37b in Appendix I). This is an equal-area projection from group (A) of polyconic projections. It is devised specifically for oblique aspect representation with the origin in latitude 45°N and longitude 5°E, so that the principal land areas of the northern hemisphere lie close to the centre of the map where distortion is smallest. The isograms shown in this figure are for maximum angular deformation (ω) at 5°, 10°, 20°, 30°, 50°, 70° and 90°. Note that the graticule shown in this figure is for 20° intervals of latitude and longitude, whereas the tabulated coordinates in Appendix II are for a 15° graticule. This choice is deliberate so that a student required to construct the projections by tables cannot get away with just tracing this graticule.

- Ordinary draughting instruments, including splines or curves and a fine needle mounted in a pin vice for pricking the positions of all points.

The procedure for plotting the intersections and drawing the graticule is as follows:

1. Select the range of coordinates which represents the maximum extent of the proposed map in each direction and determine the distance in millimetres corresponding to this range.
2. Choose an origin on the master grid which will permit the whole graticule to be plotted upon it, and number the grid lines in millimetres from the point chosen as origin using the familiar sign convention described on p. 29.
3. Extract each pair of coordinates from the table and convert these into millimetres in the master grid at the scale of the intended map. Repeat this procedure for every graticule intersection to complete a list of all coordinates converted into metric units.
4. Plot these coordinates within the master grid to locate each graticule intersection.
5. Lay a spline or curve through the succession of plotted points corresponding to the single meridian or parallel and draw a smooth curve through the points. This stage is repeated until the whole graticule has been drawn.

We shall now examine each of these steps in detail.

The use of projection tables

In this chapter we consider only the plotting of small-scale maps of the sphere. Consequently the tables are much simpler to use than those prepared for topographical cartography, which are generally based upon the equations for mapping the spheroidal surface. Such tables are briefly mentioned in Chapter 16, pp. 360–363.

Inspection of the table of rectangular coordinates for Briesemeister's projection (p. 442) indicates that the range of tabulated longitude is from the North Pole to the South Pole, but the range of tabulated longitude is only from the central meridian to 165°W. This is because the projection is symmetrical about the central meridian. Consequently the graticule intersections to the west of the central meridian can be obtained from the table by making the appropriate adjustment of the numerical value for longitude and changing the sign of x. For example, the point $\varphi = 60°\text{N}$, $\lambda = 45°\text{E}$ has coordinates:

$$x = +0{\cdot}23933, \quad y = +0{\cdot}33204$$

The coordinates of the corresponding point in the western hemisphere are for $\varphi = 60°\text{N}$, $\lambda = 30°\text{W}$, the difference in longitude resulting from the fact that the central meridian is not that of Greenwich, but 15°E. The coordinates for this point are:

$$x = -0{\cdot}23933, \quad y = +0{\cdot}33204.$$

We also see from the table that the range of the coordinate values is from x = 0·00000, y = 0·00000 for the point $\varphi = 45°\text{N}$, $\lambda = 15°\text{E}$ to the following extreme values:

Northern limit of map: $\varphi = 45°\text{N}$, $\lambda = 165°\text{W}$
$x = 0{\cdot}00000, \quad y = 1{\cdot}51188$

Southern limit of map: $\phi = 45°\text{S}$, $\lambda = 15°\text{E}$
$x = 0{\cdot}00000, \quad y = -1{\cdot}51188$

Eastern limit of map: $\varphi = 45°\text{S}$, $\lambda = 165°\text{W}$
$x = 2{\cdot}64579, \quad y = 0{\cdot}00000$

Western limit of map: $\varphi = 45°\text{S}$, $\lambda = 165°\text{W}$
$x = -2{\cdot}64579, \quad y = 0{\cdot}00000$

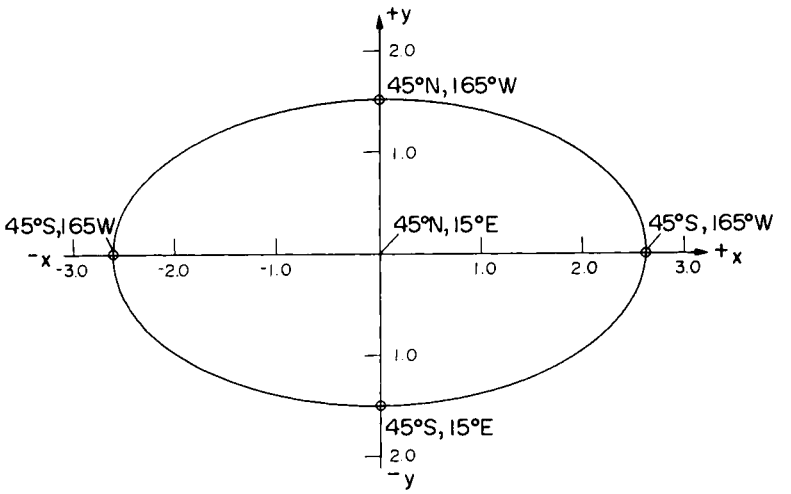


FIG. 8.03 Initial specification of the coordinates for the extreme points needed to construct a world map on Briesemeister's projection. The axes are divided in units of earth radius, R, and the four points defining the major and minor axes of the bounding ellipse have been plotted directly from the coordinates tabulated in Appendix II.

The rough draft

At this stage it is useful to make a rough plot of some of the points (e.g. the 45° graticule intersections) on graph paper using some convenient scale such as 100 mm = 1 tabulated unit, which corresponds to a representative fraction of 1/63 711 000. This is useful to find out how the tables have been compiled; for example to ascertain which direction is denoted by x. Moreover, a rough plot of this sort indicates immediately how much of the world map can be plotted from the tabulated coordinates and how much has to be plotted using different signs for either x or y. The diagram serves as a check against making gross errors in location for the first few points which are plotted on the master grid. As the work proceeds, and the pattern of points emerges, it becomes increasingly obvious where each point has to be plotted. From the extreme values tabulated above it can be seen that the world map will extend $2 \times 1{\cdot}5118 = 3{\cdot}0237$ units in the direction of the ordinate, and $2 \times 2{\cdot}6458 = 5{\cdot}2916$ in the direction of the abscissa.

Scale conversion of the tabulated coordinates

The values of x and y are given in units of earth radius. In other words, if we put R = 1·0 cm, the width of the map at that scale would be 5·29 cm.

In order to use the tabulated coordinates to plot a map at a required scale it is necessary to convert from units of R into the values of r, which, as we saw in Chapter 4, corresponds to the radius of the generating globe whose scale is the principal scale of the map. For example, it is required to plot Briesemeister's projection at 1/40 000 000. From equation (5.06), p. 82:

$$1/40\,000\,000 = r/6\,371\,100$$

and

$$r = 0{\cdot}1592 \text{ m} = 159{\cdot}2 \text{ mm}$$

This quantity is now used as a constant multiplier to convert all the tabulated values of x and y into millimetres at the scale of plotting. Thus,

$$x' = r \,.\, x \tag{8.01}$$

$$y' = r \,.\, y \tag{8.02}$$

which is an application of the scale transformation applied to cartesian coordinates described in (2.15) and (2.16) on p. 39. We have already described the (x′, y′) system in this context as the master grid coordinates.

Thus the transformed values for the four extreme points of Briesemeister's projection are:

φ	λ	x′ (mm)	y′ (mm)
45°N	165°W	0·0	+240·8
45°S	15°E	0·0	−240·8
45°S	165°W	+421·4	0·0
45°S	165°W	−421·4	0·0

This indicates that the map requires a master grid with dimensions greater than 0·843 m × 0·482 m in order to plot the world map at the required scale of 1/40 000 000.

To avoid making the calculation of the constant multiplier, r, we give these values in Table 8.02 for many of the commonly used map scales within the range 1/5 000 000 through 1/250 000 000. In addition we include the representative fractions for which r is an integer within the range 30 mm through 70 mm and for r = 100 mm.

Most pocket calculators have at least one store into which the appropriate value for r may be inserted. At the simplest level of calculation, where it is only necessary to apply the scale conversion to existing tabulated coordinates, this kind of calculator will suffice.

Construction of the master grid

We have already seen that certain classes of map projections such as the azimuthal, conical, pseudoazimuthal and pseudoconical projections of Groups C and D are more conveniently defined as functions of polar

TABLE 8.02 *Values for the radius of the generating globe,* r, *to be used as a constant multiplier for conversion of coordinates in projection tables to plot the coordinates in millimetres. This table employs the spherical assumption for* R = *6371·1 km*

Scale	r (mm)	Scale	r (mm)
1/250 000 000	25·484	1/60 000 000	106·185
1/212 366 666	30·00	1/50 000 000	127·422
1/200 000 000	31·865	1/40 000 000	159·278
1/159 277 500	40·00	1/30 000 000	212·370
1/150 000 000	42·474	1/25 000 000	254·844
1/127 422 000	50·00	1/20 000 000	318·555
1/125 000 000	50·969	1/15 000 000	424·470
1/106 185 000	60·00	1/12 500 000	509·688
1/100 000 000	63·711	1/10 000 000	637·110
1/91 015 714	70·00	1/9 000 000	707·900
1/90 000 000	70·790	1/8 000 000	796·388
1/80 000 000	79·939	1/7 000 000	910·157
1/70 000 000	91·016	1/6 000 000	1061·850
1/63 711 000	100·00	1/5 000 000	1274·220

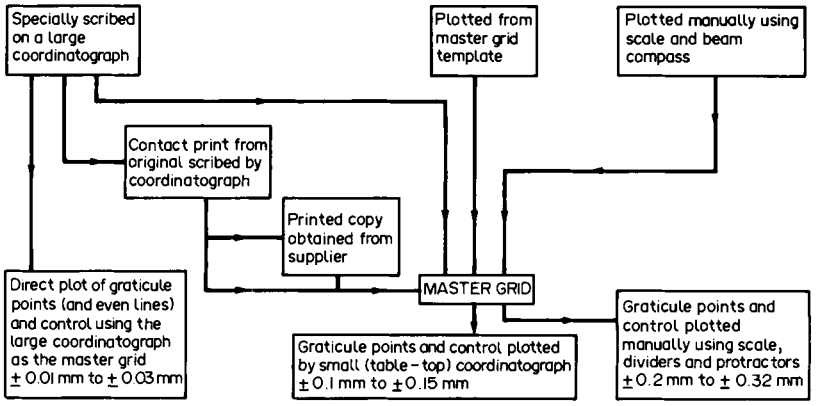


FIG. 8.04 The different methods of preparing a master grid in order to plot the graticule of a new map and any control points referring to it. (Source: Maling, 1989.)

coordinates. However, there are no polar coordinatographs available which compare in working range with large cartesian coordinatographs. Consequently it would be necessary to construct a polar master grid graphically, and this is more difficult to do than to plot a master grid in rectangular coordinates. Hence we find that *the master grid is always a system of rectangular coordinates and even when a map projection has been initially derived in polar coordinates, these are transformed into rectangular coordinates for purposes of plotting.* This is done with equations (2.01) and (2.02).

The plotting stage requires a sheet of dimensionally stable draughting film (probably polyester plastic) with dimensions greater than the maximum extent of the intended map. The preliminary drawings and computations should be consulted for these measurements. Unless a coordinatograph is used to plot the graticular intersections, a precise (x′, y′) grid must be drawn or reproduced upon the plotting sheet first. These are, as illustrated in Fig. 8.04, in preferred order of choice:

- to use a large coordinatograph as the master grid;
- to use a master grid template and small coordinatograph for plotting individual graticule intersections;
- to reproduce the master grid from one obtained by the first method, or, to use a preprinted precision grid printed on plastic;
- to construct the master grid graphically.

Use of a coordinatograph

In many cartographic establishments the whole of the work of the plotting and drawing stages of constructing a map projection can be done by

coordinatograph (Plate 1). This instrument creates the two orthogonal axes of rectangular coordinates by means of one fixed steel beam and a travelling steel gantry which has a moving plotting head attached to it. Linear displacements of the gantry and the plotting head may be transferred to scales by means of lead-screws or a rack-and-pinion movement. With the aid of verniers or micrometers attached to each movement it is possible to read or plot coordinates to a least count of 0·1 mm on the majority of instruments. Some of them even give scale readings to a least count of 0·01 mm. If the fixed beam is regarded as the y-axis, values of the ordinate may be obtained by moving the gantry along this axis to the appropriate scale reading. Values of the abscissa are changed by moving the plotting-head along the gantry. The precision of plotting is usually claimed to be a standard error of about $\pm 0{\cdot}05$ mm on each axis.

The great advantage of using a large coordinatograph to plot a map projection is that no preliminary constructions are required. A virgin sheet of plastic can be mounted on the drawing table, and may be left there until all of the graticule intersections have been plotted. Moreover, if it is required to plot a grid composed entirely of orthogonal straight lines, the fair drawing of the component lines can be done entirely on the instrument, using a special pen or scribing tool which can be fitted to the plotting head. This eliminates a great deal of slow careful work such as the alignment of a straight edge through pairs of points, which would otherwise be necessary if the lines were drawn by conventional methods.

The addition of servomotors and electronic control to a coordinatograph further extends the efficiency of the equipment, because it effectively becomes a peripheral to a computer, and can plot information automatically in either on-line or off-line mode of operation. Where the equipment has been specifically designed as computing hardware it is generally called a *graph plotter*. In addition to the obvious process of setting the plotting head to occupy a succession of calculated coordinates and plotting these, a variety of interpolation programs have been written to control the movements of the plotter as it draws or scribes smooth curves through the plotted points. Obviously this is more sophisticated than merely joining the graticule intersections by means of straight lines as if we were joining them by ruler. The earlier graph plotters (and some cheap versions still in production) used to produce lines oblique to the axes in small increments of x and y so that these had a characteristically jagged pattern. The same can still be seen on the monitor displays of some microcomputers using the cheaper kinds of graphic software. By the late 1960s the increasing sophistication of both hardware and software made possible the plotting of fine lines which appear to be free of all jagged outlines. This made possible the production of complicated graticules and also other types of lines such as the hyperbolae which have to be shown on *lattice charts* (p. 291).

The only objection to the use of a coordinatograph for plotting the graticule manually is that a large format precision instrument with a working range of about 1 m on each axis is expensive. Consequently not every cartographic establishment has access to an instrument. Therefore we must suggest some cheaper ways of obtaining the same result.

Master grid template and small coordinatograph

A master grid template (Plate 2) is a flat sheet of metal with dimensions 1 m × 0·7 m or thereabouts. This sheet has been drilled with a network of holes at uniform spacing, usually 50 mm × 50 mm or 100 mm × 100 mm. All holes are identical, and a special tool which fits them exactly is used to mark points by pricking the surface of the plotting sheet lying under the template. Although the equipment seems crude in comparison with a large-format coordinatograph, the master grid is a precision instrument, and those points which can be located with an accuracy equivalent to those plotted by coordinatograph. The job of plotting grid intersections by master grid is extremely quick, for there are no scales to be read or set. Consequently the 70 or more points drilled in a typical template can be transferred to the plotting sheet in no more time than it takes to set the coordinatograph to plot half a dozen of them. The only disadvantage of the method is that the grid points are plotted rather far apart. This means that further subdivision of the master grid by graphical methods may be needed before the required graticule intersection can be located with sufficient precision. However, a careful draughtsman who is willing to make a few additional calculations during the work of plotting ought to be able to work within the 50 mm or 100 mm grid as precisely and efficiently as within the 5 mm grid suggested earlier.

Plotting of graticule intersections can be done entirely with ordinary drawing instruments, such as spring-bow dividers and a steel scale. There is also a variety of small-size coordinatographs which can be used with the master grid template to make it practically as efficient as a large-format coordinatograph. The small coordinatograph usually has an operating range of 200 mm or less along each axis, and therefore corresponds to a miniature version of the big instrument (Plate 3). It is placed upon the surface of the plotting sheet and oriented to the points which have already been located by master grid template. The combination of the template and small coordinatograph is both efficient and cheap.

Use of a preprinted grid

We use a preprinted grid every time we plot on graph paper. However, the typical sheet of graph paper is not particularly accurately printed, and it has been reproduced upon the dimensionally unstable base of

cartridge or detail paper. For cartographic use, as a substitute to either of the instrumental methods, we need a precision grid reproduced on polyester plastic. These can be bought from some of the manufacturers of drawing office equipment, but such grids have to be tested carefully before use (see p. 168) for there are a number of indifferent products available. In a department where different projections have to be constructed fairly often, the quickest and least expensive method of producing master grids is by photomechanical reproduction of positive copies made from a precision grid which is kept solely for use as a master copy. This could be scribed by coordinatograph to the department's own specification.

Graphical construction of the master grid

In an ill-equipped drawing office, or under special working conditions, such as at sea or when the gridded sheet must be larger than the coordinatograph table, it may be necessary to construct the master grid graphically. Although we believe that the use of a preprinted grid is the more economical method to use in practice, we describe two methods of making the graphical construction. This is because useful lessons can be learnt from comparison of the two methods. One provides valuable independent checks whereas the other does not, and this important principle can be applied to the comparative study of other kinds of graphical work.

Method I (Figs 8.05, 8.06 and 8.07)

1. The approximate centre of the plotting sheet is located by drawing diagonals through the sheet corners. From the centre O, thus defined, the axes AB and CD are drawn at right angles to one another and approximately parallel to the edges of the plotting sheet. The construction of the right angle at the centre is important, for if these axes are not perpendicular the whole grid will turn out to be a parallelogram and not a rectangle.
2. The distances along the axes to the edges of the grid are set upon two beam compasses. For example, $OB = OA = 500{\cdot}0$ mm and $OC = OD = 350{\cdot}0$ mm are the settings needed to plot a grid with overall dimensions $1{\cdot}0\ \text{m} \times 0{\cdot}7$ m. The use of two beam compasses saves having to reset the measurements during subsequent stages in the construction.
3. The beam compass with the setting OC is centred at O and the arcs OC and OD are constructed on one axis.
4. The beam compass with setting OB is centred at O and the arcs OA and OB are constructed on the other axis.

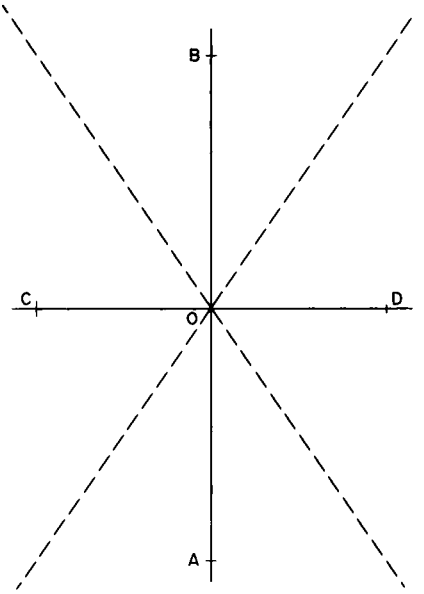


FIG. 8.05 Graphical construction of a master grid by Method I. Stages 1–4 showing the location of the centre of the plotting sheet by drawing diagonals, the construction of the axes *AOB* and *COD* and the location of the points *A*, *B*, *C* and *D* by arcs drawn from *O*.

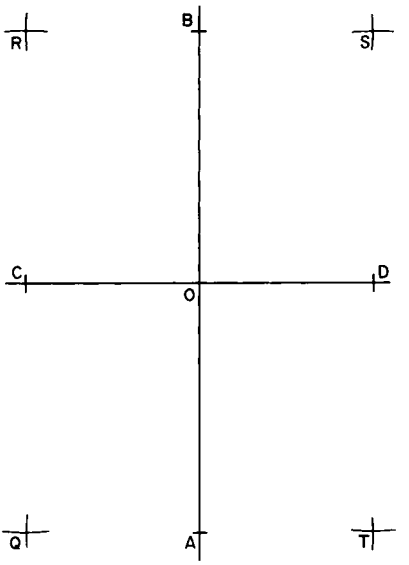


FIG. 8.06 Graphical construction of a master grid by Method I. Stages 5–8, showing the location of the corner points *Q*, *R*, *S* and *T* by the intersection of arcs drawn from *A*, *B*, *C* and *D*.

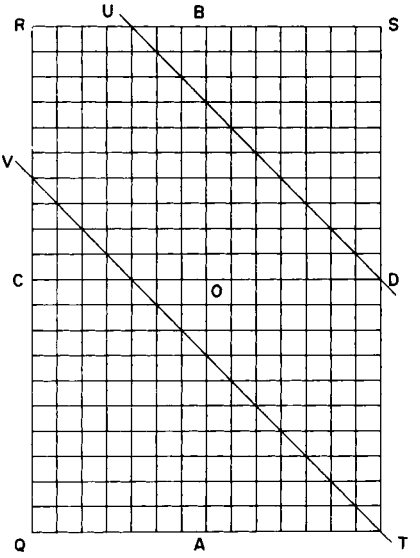


FIG. 8.07 Checking the accuracy of plotting and drawing a master grid though points such as *UD* and *VT*. Note that all grid intersections along the lines should coincide with the ruling edge.

5. The beam compass with the setting *OB* is centred at *C* to construct the arcs *CQ* and *CR*.
6. The beam compass with the setting *OB* is centred at *D* to construct the arcs *DS* and *DT*.
7. The beam compass with the setting *OC* is centred at *A* to construct the arcs *AQ* and *AT*.
8. The beam compass with the setting *OC* is centred at *B* to construct the arcs *BR* and *BS*. The intersections of arcs at *Q*, *R*, *S* and *T* locate the four corners of the grid. At this stage it is desirable to check that the length of the diagonal $QS = RT$. This is the necessary geometrical requirements for a rectangle.
9. The grid is now subdivided as required, e.g. into 5 mm units, along each side. This has to be done by setting the appropriate measurements along the beam compass and plotting each subdivision from the two most convenient control points of the eight (*A*, *B*, *C*, *D*, *Q*, *R*, *S*, *T*) which have already been located. It is most undesirable to use the drawing office methods of subdividing a line by parallel ruler, set squares or stepping off equal subdivisions by spring-bow dividers set to a separation of 5 mm. Each of these methods can introduce systematic errors into the construction and, by definition, the master grid should be sensibly free from error. See Maling in ICA (1984) for a fuller account of the technique, and Maling (1989) for an

investigation into the precision of the work. Location of a large number of subdivisions by beam compass is extremely tedious.

10. Corresponding points along the edges of the grid are joined by ruling straight lines between them.
11. The accuracy of the work may be tested by laying a straight edge diagonally across the grid, e.g. between the points D and U, T and V, etc. If all the grid intersections along that diagonal coincide with the straight edge, the construction may be accepted.

The weakness of the method is that no check is made upon the quality of the work after stage (8), when the diagonals are measured, until after the grid has been subdivided. Since stage (9) is the most laborious part of the whole job, much time and effort has been wasted if the grid proves to be unacceptable.

Method II (Figs 8.08 and 8.09)

1. The first step is to calculate the length of the diagonal of the grid ($QS = RT$) and the bearing which one diagonal ought to make with a side of the plotting sheet. This is done by plane trigonometry. For example, in Fig. 8.08

$$\tan \theta = x/y \tag{8.03}$$

$$QS = x \,.\, \text{cosec}\, \theta \tag{8.04}$$

Thus, for a grid with dimensions 1000 mm × 700 mm,

$$\tan \theta = 1000/700 = 1{\cdot}42857$$

$$\theta = 55°{\cdot}008$$

$$QS = 1000 \times 1{\cdot}22066$$

$$= 1220{\cdot}66 \text{ mm}$$

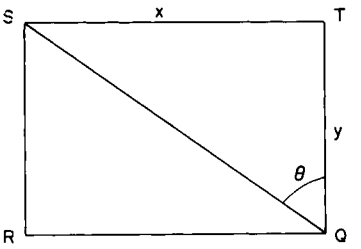


FIG. 8.08 Graphical construction of a master grid by Method II. Stage 1, showing the definition of the angle θ and determination of the length of the diagonal QS.

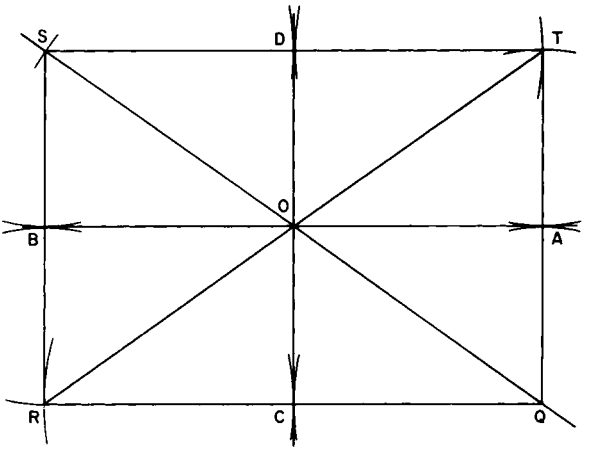


FIG. 8.09 Graphical construction of a master grid by Method II at the completion of stage 9. This shows all the arcs needed to locate the four corner points Q, R, S and T and the midpoints A, B, C and D of each side. Note that the four lines which pass through the point O all intersect at this point.

2. We commence construction from an arbitrarily chosen point near one corner of the plotting sheet we decided to call the point Q. A line corresponding to the diagonal QS is drawn from this point, making the approximate angle θ with the shorter side of the grid. This is to ensure that the sides of the grid will be more or less parallel to the edges of the plotting sheet when the construction has been completed.
3. With a beam compass set to the calculated length of the diagonal and centred at Q, construct the arc QS on the diagonal which has been drawn. This locates the point S.
4. With a beam compass set to the distance $ST = QR$, construct two arcs in the vicinity of the two remaining corners from Q and S respectively.
5. With a beam compass set to the distance $QT = RS$, construct two arcs to cut those already constructed in stage (4) from Q and S respectively. The intersection of the two pairs of arcs from Q and S locates the points T and R.
6. Using the beam compass still set to the length of the diagonal, test that $QS = RT$. If this comparison is exact, the four points define the corners of a rectangular grid. If one diagonal is longer than the other, the figure is a parallelogram and the construction must be repeated.
7. Join RT.
8. Join the four corners of the grid and bisect each side. This defines a further four points, A, B, C and D.
9. Join AB and CD. If the four lines AB, CD, QS and RT all meet at the point O the construction is correct. Any errors in construction

are indicated by a *cocked hat*, which is the triangle formed by lines which fail to pass through the centre. This must be eliminated by repeating the construction.

10. Subdivision of the grid is done in the same way as stage (9) of Method I.
11. Testing of the final construction is carried out in the same way as stage (10) of Method I.

The advantage of Method II is that this contains two independent checks upon the quality of the work before the tedious job of subdividing the grid is attempted. This means that comparatively little time has been wasted if the first few attempts fail to produce a grid of sufficient standard.

Drawing the graticule

We need not comment in detail about plotting within the master grid, apart from noting that this is most easily done by linear measurements, using a spring-bow or similar dividers and plotting each point by means of four measurements made from each corner of the grid square in which the point is located. Only two of these measurements are needed to locate a point, but four are used to overcome the possibility of both gross and systematic errors in plotting. Graticule intersections are plotted on topographical maps by the *arc and tangent method* described in US Army (1955) and Ministry of Defence (1962).

Where the draughting specification calls for parallels and meridians to be drawn in full, and they are curved, it is necessary to use *scale-assisted* draughting methods using curves or splines as the aids to draughting. The curves representing the parallels and meridians on a map are lines which satisfy specific mathematical functions, and these functions must be satisfied not only at the graticule intersections which have been plotted but at all intervening points along each curve. Hence the smooth curve joining the plotted points has mathematical significance and it will not suffice to draw it in any arbitrary fashion. Bearing in mind that the standard error in location of the graticule intersections ought to be about $\pm 0{\cdot}1$ mm, this suggests that considerable care must be taken in selecting how each line passes through the points.

One important aid to construction is to plot more graticule intersections than need to be shown on the completed map. For example, if the final map is to show 20° parallels and meridians it is worth plotting them at least at 10° intervals, or even for every 5° on the plotting sheet. Although this involves a massive increase in the amount of computation and plotting time required, it has additional value as an aid to any graphical compilation work, because the closer graticule affords more and better control in transferring map detail from one scale and projection to

another. Of course these recommendations have been largely superseded by the developments in computer graphics. If this work can be done by computer on a graph plotter, the graphical difficulties disappear.

It is difficult to lay down any formal rules concerning the use of splines and curves for completing drawing of the graticule. For illustrations of the manipulation of the various drawing instruments and aids which may be used, see both Maling and Kanazawa in ICA (1984). The difficulty of this part of the work depends very much upon the complexity of the lines which have to be drawn. Clearly a projection comprising straight lines and circular arcs is much easier to draw than one like Briesemeister's projection illustrated in Fig. 8.02. However the following advice may be useful.

Curves (Plate 4)

The procedure is to test different parts of different curves to find which of them best suits the plotted distribution of points. It is not sufficient to find a curve which appears to pass smoothly through two or three intersections. The ruling edge of the curve must pass through about four or five consecutive points in order to draw only a short part of the required line. The reason for this is illustrated in Fig. 8.10. In order to draw a smooth curve through the plotted points *a*, *b*, *c*, *d*, *e*, *f* and *g* it will be necessary to fit various curves in different positions. To draw the line

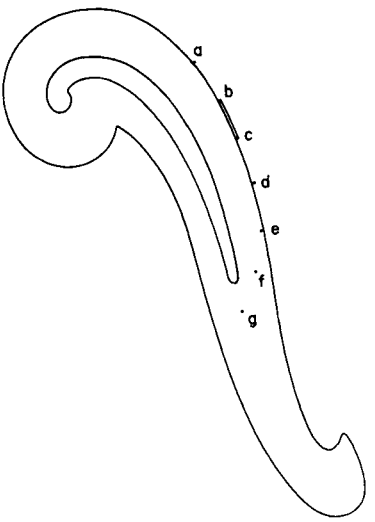


FIG. 8.10 The alignment of a French curve through four plotted points in order to draw the portion *bc* of the curve.

it may be necessary to fit a curve to the points *a*, *b*, *c* and *d* in order to draw the portion *bc*; to fit it through *b*, *c*, *d* and *e* to draw *cd* and so on. If a ruling edge can be found to fit the points *a*, *b*, *c* and *d* in one position (as illustrated in Fig. 8.09), followed by a setting through *d*, *e*, *f* and *g* in a second position, it is likely that the two lines drawn to meet at *d* would result in an unintended discontinuity at that point.

A further difficulty is that the ruling pen or scribing tool held in the optimum position for drawing is slightly offset from the centre of the curve. Consequently the ruling edge of the curve has to be slightly offset from the points, so that the nib or sapphire passes through each point correctly. Thus, in addition to trying to make the curve fit a sequence of points, it is also desirable to imagine it tracing a line which is parallel to that required. All this calls for considerable skill.

Splines (see Plates 5 and 6)

The draughtsman's spline is a flexible rod about 1 m in length. Traditionally this was made of lance-wood, though nowadays a variety of other flexible materials can be used. The rod may be square or rectangular in section. It may taper towards the ends or towards the middle. Only trial and error shows how much curvature can be obtained from a rod of particular length and cross-section, and therefore how many weights will be needed to hold it in position. The location of the ship weights is important from the point of view of stability and continuity of drawing. It is undesirable to have weights holding the spline in position along that edge where the curve has to be drawn. Every time the pen approaches a weight, drawing must be interrupted and a small gap has to be left in the line. The interruptions have to be made good later, and accurate matching across the gaps is exceptionally difficult. Usually it is better to have the weights aligned on the inside (concave) surface of the spline and to draw along its outer or convex edge, but ultimately there must be at least two weights on the outer edge to hold the spline in position (Plate 5). Stability in the position of the ruling edge is important, for it is obviously unsatisfactory if the spline yields to the slight lateral pressure as the pen moves along its side. In ink draughting this invariably causes smudging of the freshly drawn line. In any case it reduces the accuracy of the drawing.

A different technique for anchoring the spline to the fair drawing may be used when the work is done by scribing. Since the opaque surface of scribe coating is unaffected by materials which would ruin an ink drawing on paper or card, we may use modelling clay, such as Plasticene, to replace some of the ship weights. The spline to be used for this job can be cut from heavy-gauge plastic, for example, Cobex of thickness 1 mm, in a strip of width 5 mm to any desired length.

This spline is temporarily held in the upright, or edge-on, position on

the manuscript by means of ship weights. The modelling clay is packed along the inside edge throughout its length as illustrated in Plate 6. For large radius curves it is sufficient to use only two ship weights at the ends of the spline; the whole of the remainder being held in position by clay. A small radius curve or a complicated curve with reverse curvature will need a few additional ship weights to keep the spline firmly in position.

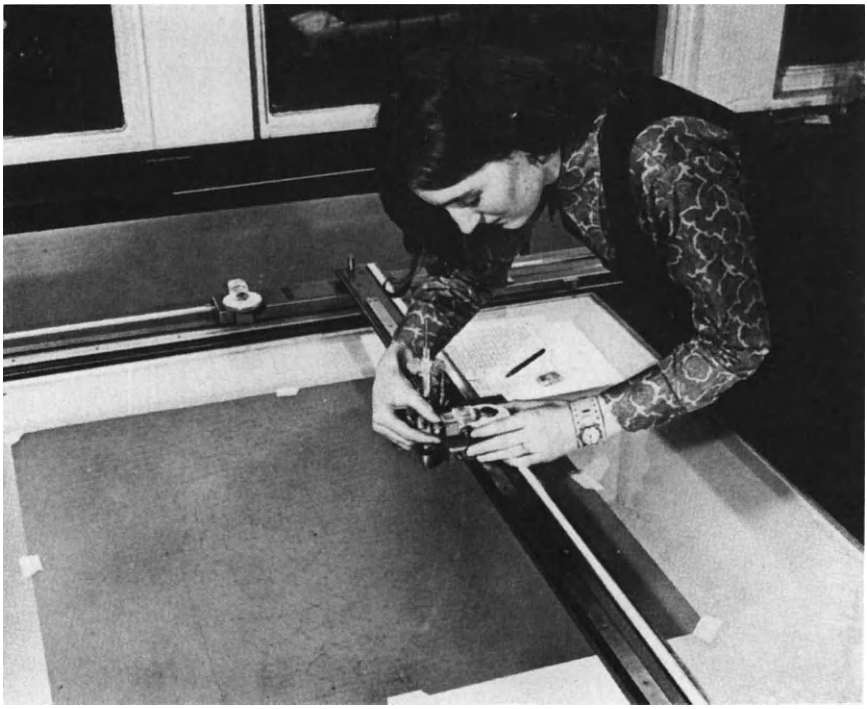


Plate 1 Coradi coordinatograph with effective plotting dimensions 1000 mm. in both *x* and *y*.

Plate 2 Haag-Streit Master Grid Template with effective plotting dimensions 1000 × 700 mm. and holes at 100 mm. intervals.

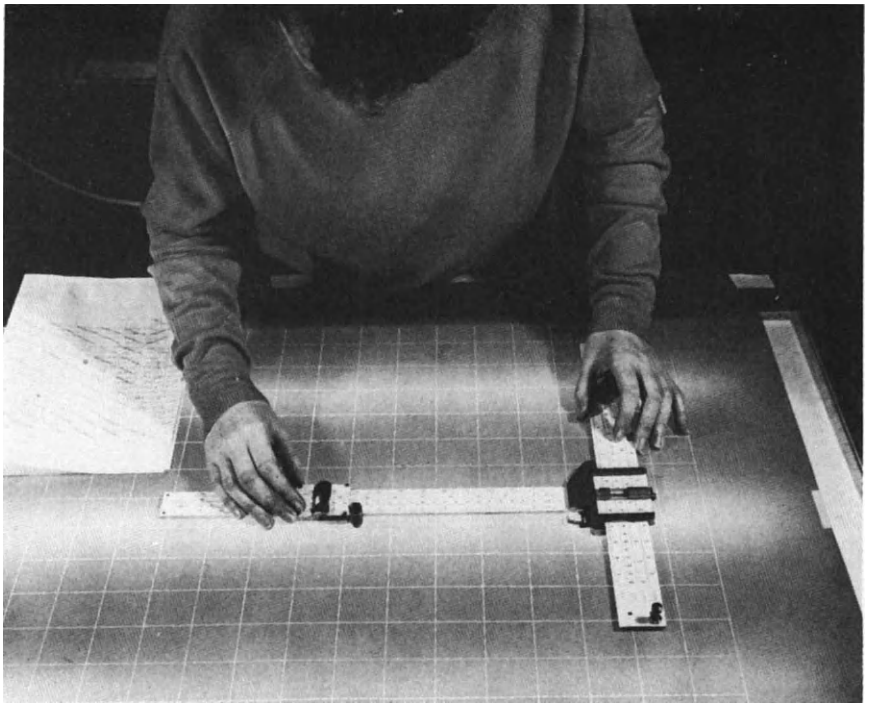


Plate 3 Aristo small size coordinatograph for use with a Master Grid Template.

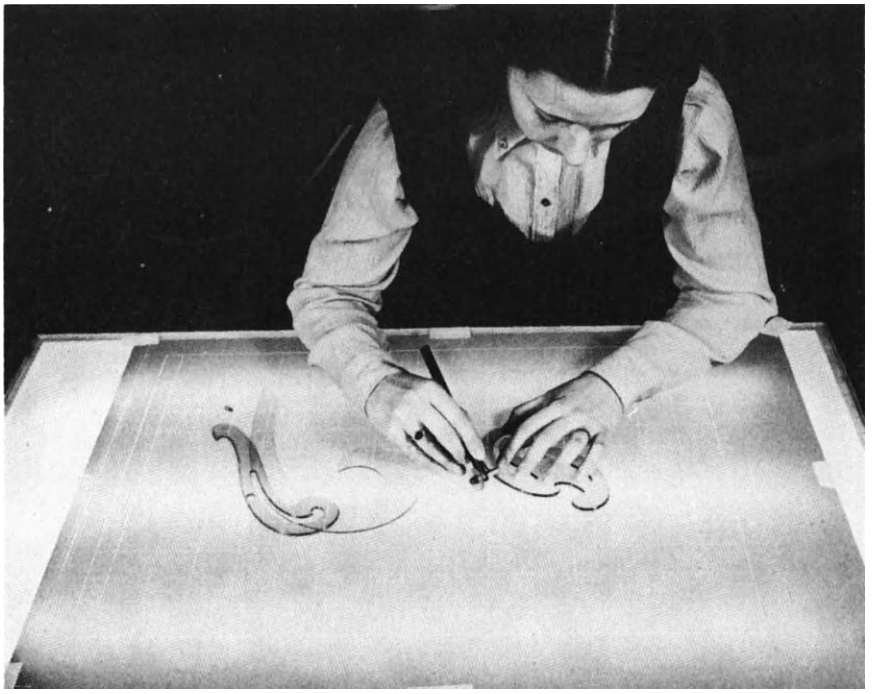


Plate 4 Scribing curved lines with the aid of a French Curve.

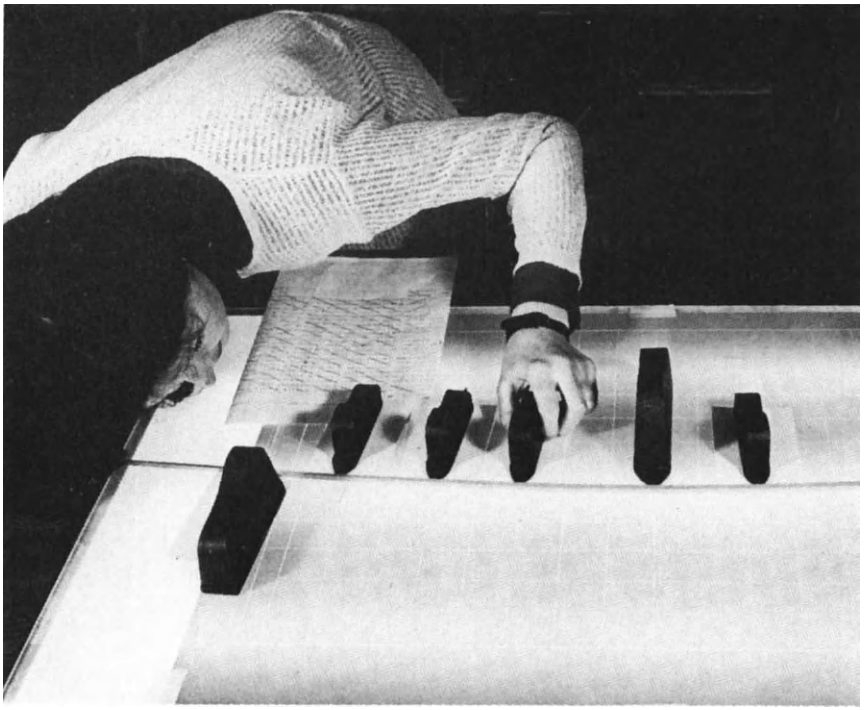


Plate 5 Alignment of a conventional lancewood spline held in position with ship weights.

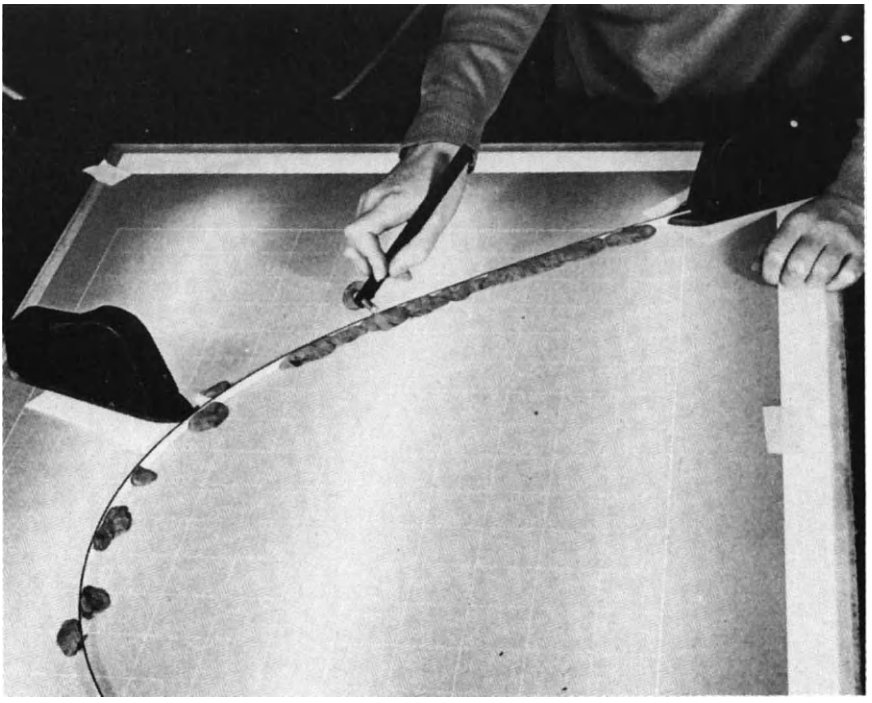


Plate 6 Scribing curved lines with the aid of a plastic spline held in position with weights

CHAPTER 9

Computation of projection coordinates

> It is unworthy of excellent men to lose hours like slaves in the labour of calculation which could safely be relegated to anyone else if machines were used.
> Leibnitz, 1671

Introduction

Since we recommend that a map projection should be constructed by plotting the plane coordinates of every graticule intersection, we must assume that the necessary tables or software are already available, or provide instruction how to produce them. Although some coordinate tables are given in many of the standard foreign textbooks, such as Driencourt and Laborde (1932), Reignier (1957), Wagner (1949), few of them have ever been published in the English language. Consequently they are much less well known than they ought to be. Space does not allow us to provide a comprehensive collection of tables in this book, apart from the example of Briesemeister's projection in Appendix II. In any case, no published set of projection tables can provide the coordinates for every oblique aspect version of every projection. Therefore it is necessary to know how to compute the appropriate coordinates.

Methods of calculating coordinates

The majority of coordinate computations which are listed in Appendix I are not difficult to solve numerically, and can be done on a pocket calculator. The repetitive nature of the work is particularly suitable for tackling the problem by microcomputer rather than by hand. Considerable economies in computing arise from exploitation of the fact that many map projections are symmetrical about one or two axes. A 10° graticule of the world comprises 634 graticule intersections, but it may be necessary only to compute the coordinates of the graticule intersections in one hemisphere (327 points), or even in one quadrant (173 points) with corresponding reduction in time. If, however, the intention is to plot some

map detail too, even if this is only a plot of coastlines on a very small-scale world map, the number of transformations to be computed increases to several tens or even hundreds of thousands of points. Then the relative economy of the method of computation becomes important.

Older methods of computation

The first edition of this book appeared in the early 1970s. At that time digital computing was still confined to batch processing on a mainframe computer through the medium of punched cards or paper tape. In order to compile or run a program it was necessary to deliver a deck of cards, or a spool of tape, to the computing centre and then await the return of the program and data, together with any output, one or two days later, usually to find that the processing had failed for some reason or other and that no results had been obtained. This was at least a decade before the term 'user-friendliness' was coined; indeed, at that time such a concept would have seemed incomprehensible, for all computing systems were distinctly unfriendly to the casual user. With hindsight it seems remarkable that much progress was ever made. The alternative was to use the conventional methods of the day, which were the mechanical calculator, supported by seven- or eight-figure tables of trigonometric functions, and using a slide rule to interpolate proportional parts between the tabulated values. This was laborious. For example, the work done by the author on the Hammer–Aitoff projection (Maling, 1962), was started in this way. Even with the help of pre-printed forms, this required about 20–30 minutes to compute the coordinates for each graticule intersection. Thus it would have required more than 300 hours of computing to produce just one asymmetrical (skew oblique or plagal) version of it. Computation of the distortion characteristics, involving the computations given in the worked example on pp. 118–121 took longer, even for a network of only 50 points. The work was completed using an IBM 1620 computer. This reduced the time needed to 15 seconds per point, 9 seconds to make the calculations followed by a pause for 6 seconds when the machine had to stop computing so that the results could be output by electric typewriter. Nowadays this seems unbelievably slow. The first pocket calculators became available in the early 1970s, about the same time that the first edition of this book was published, but the desk-top microcomputer did not appear in large numbers or at reasonable cost until 5 or 6 years later. It follows that the chapter in the first edition corresponding to this now has a distinctly old-fashioned flavour; for it deals with such subjects as the relative merits of solving spherical triangles by logarithms and by machine, and using the various haversine formulae rather than the standard equations of spherical trigonometry in order to remove ambiguity in the results. Those who still need to know about these methods are

referred to the first edition of the book; they will not find that information here, for the data-processing revolution which has occurred in the past 15 years has swept them all away as practical applications.

Modern computing methods

The kind of computations to be described ought not to present any difficulties to the reader who has learnt to use a 'scientific' pocket calculator or, even better, can write simple programs for a microcomputer using FORTRAN, BASIC, PASCAL, or one of the other fashionable programming languages. We do not offer any program listings here. A feature of the accelerating progress in data-processing is that new operating systems and programming languages soon render existing methods obsolete. Many of the programs published in textbooks 10 or 15 years ago refer to machines with operating systems which have long been superseded. The conversion of these into a form compatible with current systems is usually fraught with difficulty, and it is generally easier to start anew than to attempt to revive old programs.

We should, however, bear in mind that the equations themselves may be rewritten with profit, agreeing with the comment by Vincenty (1971) that:

> In order to utilize an electronic desk computer to the fullest extent, efficient programs must be written for it. This, in turn, means that many existing formulas designed for use with logarithms, rotary calculators and tables must be rewritten in a form which suits the machine best. Many seemingly impossible programs can be written for a relatively inexpensive desk machine if more thought is given to recasting the equations than to actual programming.

Vincenty then proceeds to provide an example which relates to the calculation of the radius of the rectifying sphere which has been described by equation (4.42), p. 78, using the notation of Adams (1921), who wrote:

$$R = a(1-n)(1-n^2)(1+9/4n^2+225/64n^4\ldots) \qquad (9.01)$$

where a is the major semi-axis of the spheroid and $n = (a-b)(a+b)$ as in Chapter 4. Vincenty (1971) has modified this equation so that R is expressed as a function of the polar radius of curvature, c, instead of the equatorial radius and the square of the second eccentricity, e'^2, rather than the first eccentricity (e^2).

Substituting $a = c(1-n)/(1+n)$ in (9.01), expanding and collecting terms we obtain:

$$A = R/c = 1-3n+21/4n^2+31/4n^3+657/64n^4 \qquad (9.02)$$

which can be written as

$$R/c = 1-1/64n(192-n(336-n(496-657n))) \qquad (9.03)$$

This gives the working equation in a 'nested' form which is very convenient for programming, as it does not require storing of intermediate values.

Typical computations needed in small-scale (atlas) cartography are to derive an oblique aspect graticule to satisfy the particular requirements for a new map. The comparable need in large-scale (topographic) cartography usually comprises the transformation of the positions of points from one version of one projection to a different version or a different projection. We treat with these in Chapters 15, 16 and 19.

Change of aspect

This usually involves three different coordinate transformations which have to be carried out consecutively for every graticule intersection.

- The transformation from geographical coordinates (φ, λ) to bearing and distance coordinates (z, α) on the sphere.
- The transformation from bearing and distance coordinate into projection coordinates (x, y) or (r, θ) on the plane.
- The transformation from projection coordinates into master grid coordinates (x′, y′) for plotting. This part of the transformations corresponds to *scale conversion of the tabulated coordinates*, described in Chapter 8, pp. 161–162.

Diagrammatically these transformations may be written:

$$(\varphi, \lambda) \rightarrow (z, \alpha) \rightarrow (x, y) \rightarrow (x', y')$$
$$\hookrightarrow (r, \theta) \rightarrow (x, y) \hookrightarrow$$

An alternative is to convert latitude and longitude into three-dimensional cartesian coordinates (X, Y, Z) on the sphere; then we rotate these axes to obtain (X*, Y*, Z*) coordinates of each point in the new aspect. These coordinates are converted back into spherical polar coordinates. Next the map projection equations are applied and finally the master grid (x′, y′) coordinates are obtained. Diagrammatically these transformations may be written:

$$(\varphi, \lambda) \rightarrow (X, Y, Z) \rightarrow (X^*, Y^*, Z^*) \rightarrow (\varphi', \lambda') \rightarrow (x, y) \rightarrow (x', y')$$

Bearing and distance coordinates

This system of spherical polar coordinates was introduced briefly in Chapter 3 as an alternative to geographical coordinates. From the practical point of view of constructing oblique aspect map projections they are a valuable aid, because the use of them generally overcomes any need

to determine and evaluate complicated algebraic expressions relating (φ, λ) to an origin (φ_0, λ_0) of the projection and thence to the (x, y) plane coordinates of the map. By introducing bearing and distance coordinates, we split the transformation from the spherical surface to the plane into two separate operations.

Consider the part of the spherical surface illustrated in Fig. 9.01. We are accustomed to define the positions of the two points A and B by means of their geographical coordinates, but we could also define them by the (χ, λ) coordinates of colatitude and longitude. The only difference between this system and conventional geographical coordinates is that colatitude (χ) is measured from the pole in a plane containing the axis of rotation rather than from the plane of the equator. In other words, we use the angle NOA. The (χ, λ) graticule differs from the conventional system of geographical (φ, λ) coordinates in only one respect. The numerical values assigned to parallels of colatitude increase outwards from the geographical pole towards the equator.

Suppose, however, that we wish to refer the point B to the point A rather than to the pole N as shown in Fig. 9.02. We may imagine an ordered sequence of small circles and great circles to which A is the pole as being the result of shifting the entire pattern of (χ, λ) coordinate lines from N to A. In this system the position of B is related to that of A by means of the angular distance $AOB = z$, measured at the centre of the sphere, together with another angle, such as NAB measured between the planes NOA and OAB. By analogy with geographical coordinates the second angle is also represented by the spherical angle NAB. If we refer

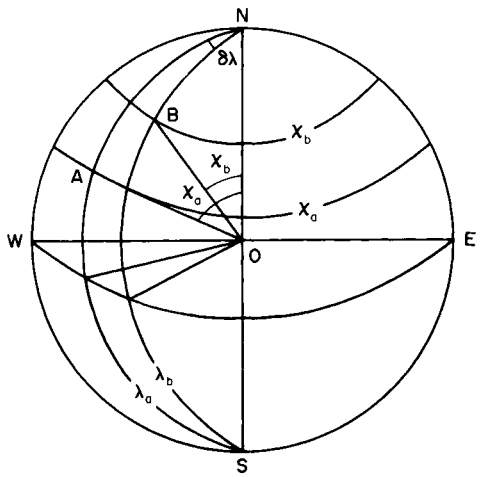


FIG. 9.01 Points A and B on the spherical surface and their definition by means of colatitude and longitude (χ, λ) coordinates.

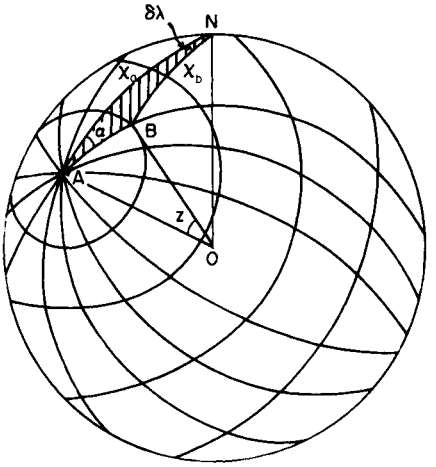


FIG. 9.02 The bearing and distance coordinates (z, α) of a point B from a pole A.

back to the spherical triangle illustrated in Fig. 3.04 on p. 55, we find that the angle NAB is the true bearing, α, of B from A. Combining these two measures to form an alternative system of coordinate reference upon the spherical surface, we have defined the (z, α) system of *bearing and distance coordinates*.

Thus we may define the position of any point on the curved surface with respect to any other point which has been selected as the pole for the (z, α) system. If this happens to be the geographical pole, then $z = \chi$ and $\alpha = \lambda$ so that the small circles representing $z =$ constant become parallels of colatitude and the circles denoting $\alpha =$ constant become meridians.

The transformation from geographical coordinates into bearing and distance coordinates is accomplished by the solution of the spherical triangle NAB. This has already been explained in Chapter 3, and it is only necessary to convert the algebraic notation into the forms which are most commonly encountered in practice. We denote the coordinates of the origin, A as (φ_0, λ_0) and those of the other point, such as B as (φ, λ). The difference in longitude between them is $\delta\lambda = \lambda_0 - \lambda$. Then by substituting these terms in equations (3.20) and (3.24).

$$\cos z = \sin \varphi_0 \,.\, \sin \varphi + \cos \varphi_0 \,.\, \cos \varphi \,.\, \cos \delta\lambda \tag{9.04}$$

$$\sin \alpha = \cos \varphi \,.\, \sin \delta\lambda \,.\, \operatorname{cosec} z \tag{9.05}$$

Hence the first transformation $(\varphi, \lambda) \rightarrow (z, \alpha)$ involves the numerical solution of equations (9.04) and (9.05).

Snyder (1987a) has noted that equation (9.04) is not particularly accur-

ate in practical computation for values of z close to zero. He suggests a rearrangement of the equation used by astronomers into the form

$$\sin \tfrac{1}{2}z = [\sin^2(\tfrac{1}{2}\delta\varphi) + \cos\varphi \,.\, \cos\varphi_0 \,.\, \sin^2(\tfrac{1}{2}\delta\lambda)]^{1/2} \qquad (9.06)$$

It is also possible to use a version of (9.04) which eliminates the term in z. After some rearrangement, we have

$$\tan\alpha = \cos\varphi_0 \,.\, \sin\delta\lambda / [\cos\varphi \,.\, \sin\varphi_0 - \sin\varphi \,.\, \cos\varphi_0 \,.\, \cos\delta\lambda] \qquad (9.07)$$

The advantage of equation (9.06) over (9.04) is that use of it avoids inaccuracies in finding the inverse sine of an angle close to 90° or the inverse cosine of an angle close to 0°.

Transformation from bearing and distance coordinates to projection coordinates and master grid coordinates

In Chapters 5, 6 and 7 we have made use of the general functions to relate the geographical coordinates of a point to its position by means of the (x, y) or (r, θ) systems. In Chapter 10 we shall derive certain map projections by analytical methods. The first of these is the *Azimuthal equal-area (Lambert) projection*, the derivation of which illustrates the use of polar coordinates. Equations (10.19), p. 200 has the form

$$\left.\begin{aligned} r &= 2 \,.\, \sin(\chi/2) \\ \theta &= \lambda \end{aligned}\right\} \qquad (9.08)$$

where χ is the colatitude, and R = 1.

The description of bearing and distance coordinates has emphasized that the (χ, λ) system is just a particular case of the (z, α) system where the point *A* is at the geographical pole. It follows that where a projection is described in terms of colatitude and longitude the transformation to bearing and distance coordinates requires no more than substitution of x for χ and α for λ. Therefore (10.19) may also be written in the form

$$\left.\begin{aligned} r &= 2 \,.\, \sin(z/2) \\ \theta &= \alpha \end{aligned}\right\} \qquad (9.09)$$

and any transverse or oblique aspect of the projection can be derived once we known the (z, α) coordinates for each graticule intersection which we wish to plot.

However a third stage of transformation is required before the map can be constructed to known scale. It was emphasised in Chapter 8 that map projections are invariably constructed on a master grid of rectangular coordinates. Hence it is required to transform the polar coordinates of (9.08) and (9.09) to a cartesian system.

In the normal aspect of the Azimuthal equal-area projection the geo-

graphical pole is the origin of the (r, θ) system of polar coordinates. Therefore we make this point the origin of the (x', y') system of master grid coordinates. We further specify that the Greenwich Meridian coincides with the $-y'$ axis, as illustrated in Fig. 9.03. Then any point whose polar coordinates have been expressed by (9.08) may be located on the master grid by the equations

$$\left.\begin{aligned} x' &= 2r \,.\, \sin(\chi/2) \,.\, \sin\lambda \\ y' &= 2r \,.\, \sin(\chi/2) \,.\, \cos\lambda \end{aligned}\right\} \tag{9.10}$$

where r is the common multiplier obtained from Table 3.02, p. 162 or by solution of equation (5.06). For example, if the point B has geographical coordinates 60°N, 30°E and the scale of the map is 1/20 000 000, so that r = 318·55 mm, the master grid coordinates of this point are

$$\begin{aligned} x' &= 2 \times 318{\cdot}55 \times \sin 15^\circ \times \sin 30^\circ \\ &= 82{\cdot}45\,\text{mm} \\ y' &= -2 \times 318{\cdot}55 \times \sin 15^\circ \times \cos 30^\circ \\ &= -142{\cdot}80\,\text{mm}. \end{aligned}$$

The position of this point is shown in Fig. 9.03.

In the oblique aspect of the Azimuthal equal-area projection illustrated by Fig. 9.04, p. 184, the origin of the polar coordinates is the point (φ_0, λ_0) which we further specify as $\varphi_0 = 40°\text{N}, \lambda_0 = 0°$. We make this the origin of the master grid coordinates and further specify that the central meridian southwards from this point coincides with the $-y'$ axis. Equation (9.07) may now be transformed into

$$\left.\begin{aligned} x' &= 2r \,.\, \sin(z/2) \,.\, \sin\alpha \\ y' &= 2r \,.\, \sin(z/2) \,.\, \cos\alpha \end{aligned}\right\} \tag{9.11}$$

This time the point B ($\varphi = 60°\text{N}, \lambda = 30°\text{E}$) must be related to the origin by its bearing and distance coordinates. By calculation or from tables, we find $z = 41°34'01''$ and $\alpha = 40°44'23''$. Therefore the master grid coordinates are

$$\begin{aligned} x' &= 2 \times 318{\cdot}55 + \sin 20°47'00'' \times \sin 40°34'23'' \\ &= 147{\cdot}54\,\text{mm} \\ y' &= 2 \times 318{\cdot}55 \times \sin 20°47'00'' \times \cos 40°44'23'' \\ &= 171{\cdot}29\,\text{mm} \end{aligned}$$

The position of this point is shown in Fig. 9.04.

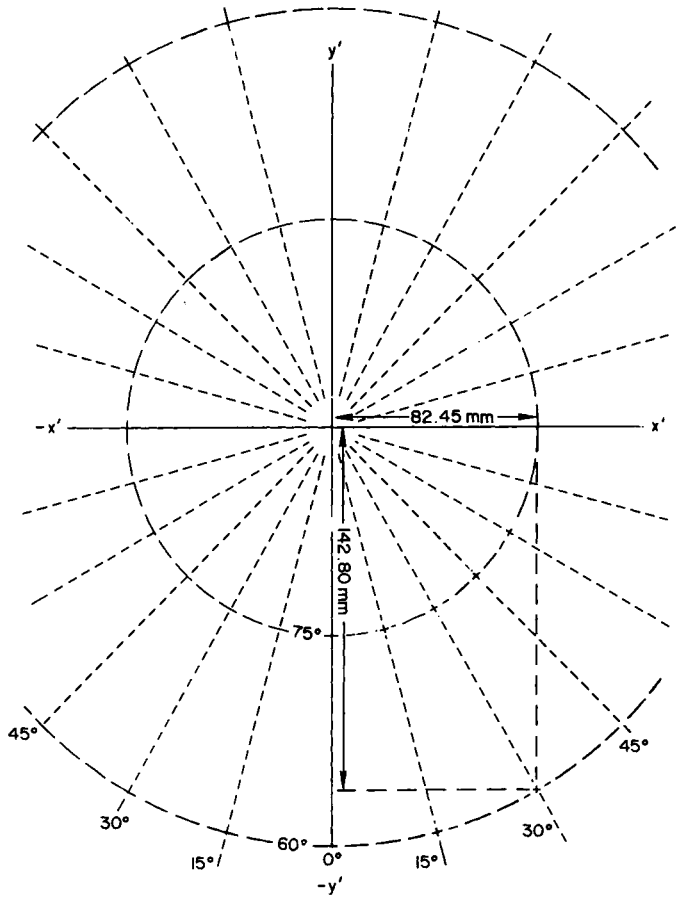


FIG. 9.03 The construction of a normal aspect Azimuthal equal-area projection by master grid coordinates.

Hammer's Tables

The (z, α) system has been the preferred method for determining the coordinates of map projections in their transverse or oblique aspects because tables for (z, α) coordinates were available for the century before we have programmable pocket calculators and microcomputers. The original tables, specifically prepared for cartographic use, were known as *Hammer's Tables* after their inventor, Professor E. Hammer. Other tables providing solutions of spherical triangles have also been produced for astronomical navigation and these, too, were easily adapted to cartographic use although, as described in the first addition, such tables are usually truncated for very small ($< 5°$) and very large ($> 80°$) angles, because it is unwise to use these extreme values of altitude for astro-

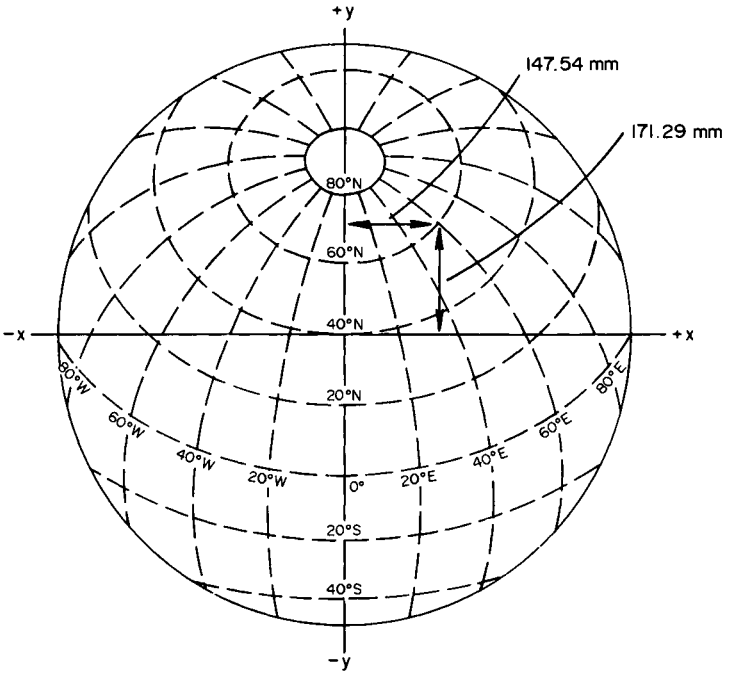


FIG. 9.04 The construction of an oblique aspect Azimuthal equal-area projection by master grid coordinates.

navigation. The first edition of this book was prepared at a time when the use of tables was normal, and the corresponding chapter in that edition contained much information about which tables were available and how they were used. However, the determination of the (z, α) coordinates, even using a pocket calculator, is such a small job that printed tables have now become virtually obsolete.

Because the (z, α) transformations is particularly easy to use with the azimuthal projections, it is the members of this class which have most commonly been presented in their transverse and oblique aspects. Indeed, many textbooks convey the impression that only the azimuthal class of projections can be presented in different aspects. This supposition is incorrect, as indicated by the illustration of different versions of the Cylindrical equal-area, Sinusoidal and Mollweide's projections in Chapter 7. However, it is still unusual to see oblique aspect pseudocylindrical and polyconic projections. Although it can be done, the application of bearing and distance coordinates to some of these classes leads to some tortuous calculations. Indeed, both Lee (1944) and Wray (1974) consider that the (z, α) transformation is only suitable for the so-called 'conical projections', which in some classification systems corresponds to the three

Group D categories of azimuthal, conical and cylindrical projections. This does not mean that the (z, α) transformation cannot be used for other categories of projection. The present author used the (z, α) system for the oblique versions of the Hammer–Aitoff projection, but thereby created a number of difficulties which might have been circumvented by using a more general approach.

Transformation through three-dimensional cartesian coordinates

The alternative to this use of the (z, α) system is to consider a three-dimensional cartesian coordinates system, with rotations about the three axes, and the rigid-body rotation of the vector formed by the radius joining the centre of a sphere to a point A on the spherical surface. The reason why this method of defining aspect was not used earlier is not that it was unknown; indeed, the geometry of mapping a sphere to itself was investigated by Cayley in the 1840s. The method was not used in practice simply because it involved some formidable computations.

Euler's angles

In Chapter 2 we introduced the geometry of the transformation from one grid into another, and recognized the three motions of translation, scale change and rotation applied to the (X, Y) or (E, N) axes of plane systems. The rotation matrix for the two-dimensional transformation was derived in equations (2.18)–(2.34). The corresponding expressions for the rotation of a three-dimensional cartesian system are now investigated through the medium of the three Eulerian angles, α, β and γ.

The starting point is a sphere whose surface may be defined by a geocentric three-dimensional cartesian system. The origin of the coordinates is the point O, which is the centre of the earth, and it is further specified that the Z-axis initially coincides with the earth's axis of rotation, so that the point Z on the spherical surface is the North Geographical Pole. This means that the X and Y axes both lie in the plane of the equator and that X and Y are separated by 90° in longitude. For example, if X corresponds to the point where the Greenwich Meridian intersects the equator, Y is the intersection of the meridian 90°E with the equator. It follows that the spherical triangle XYZ is *trirectangular*, having three angles which are all right angles and three sides which are also of length 90°. It therefore represents one-eighth of the total spherical surface. We need to investigate what happens if we rotate the coordinate system about one or all the axes.

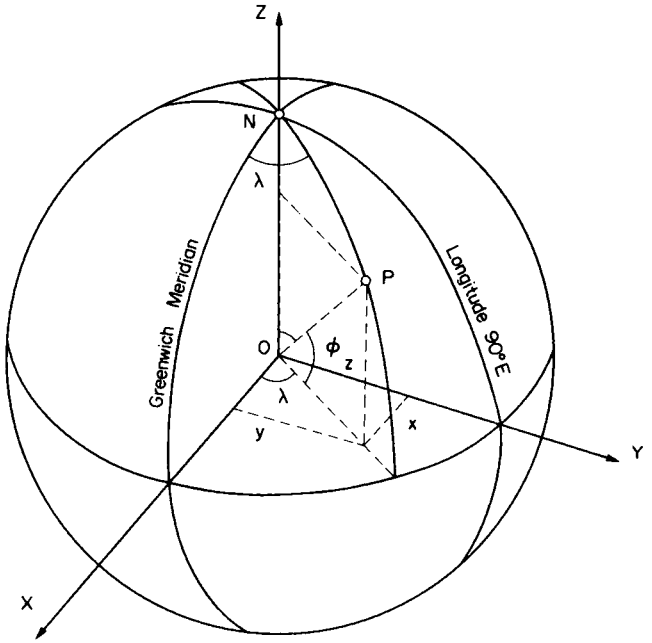


FIG. 9.05 The definition of the spherical surface by means of three-dimensional cartesian coordinates.

Rotation α about the Z-axis

Consider the two-dimensional figure corresponding to the plane of the equator. If we apply a rotation α about the Z-axis, this gives rise to a shift in the positions of X and Y which now occupy the points X′ and Y′.

Written in full, the coordinates of the new position of the axes (X′, Y′, Z′) are

$$X' = X \,.\, \cos\alpha + Y \,.\, \sin\alpha \tag{9.12}$$

$$Y' = -X \,.\, \sin\alpha + Y \,.\, \cos\alpha \tag{9.13}$$

$$Z' = Z \tag{9.14}$$

Because this rotation is about the Z-axis we see that the value of Z has no effect upon either the $X \rightarrow X'$ or the $Y \rightarrow Y'$ transformations. Moreover, the rotation about Z has no affect upon its own position. Thus, in matrix notation we may write

$$\mathbf{R}_Z = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{9.15}$$

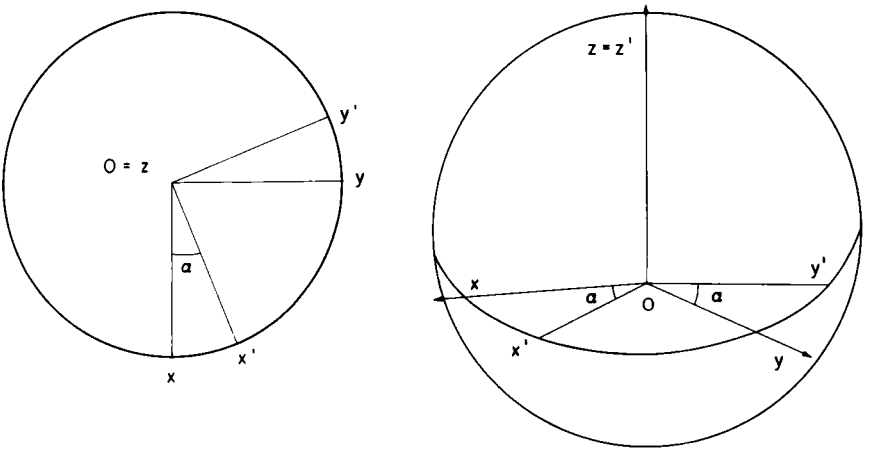


FIG. 9.06 Definition of the Eulerian angle α.

and we may express the three equations (9.12), (9.13) and (9.14) as

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \mathbf{R}_Z \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{9.16}$$

Rotation of γ about the Y′-axis

The next stage is to consider the rotation γ about the Y′-axis. From Fig. 9.07, the rotation matrix is

$$\mathbf{R}_Y = \begin{pmatrix} \cos\gamma & 0 & \sin\gamma \\ 0 & 1 & 0 \\ -\sin\gamma & 0 & \cos\gamma \end{pmatrix} \tag{9.17}$$

so that

$$\begin{pmatrix} X'' \\ Y'' \\ Z'' \end{pmatrix} = \mathbf{R}_Y \cdot \begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} \tag{9.18}$$

$$= \mathbf{R}_Z \cdot \mathbf{R}_Y \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{9.19}$$

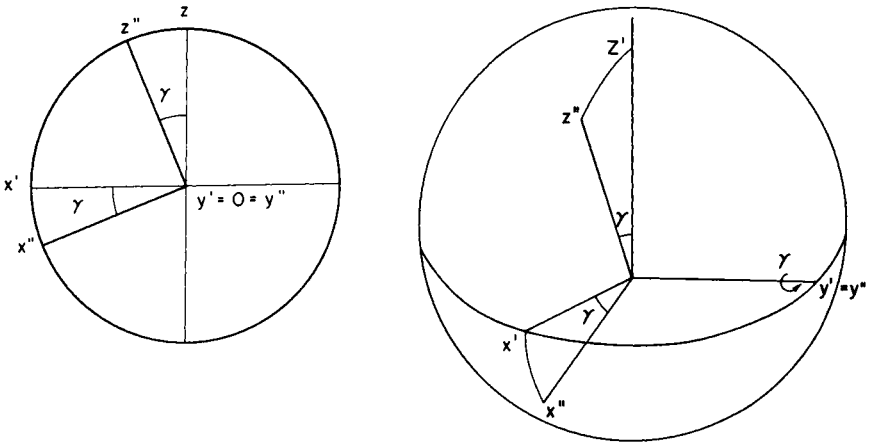


FIG. 9.07 Definition of the Eulerian angle γ.

The equation relating the coordinates before and after rotation is

$$\begin{pmatrix} X'' \\ Y'' \\ Z'' \end{pmatrix} = \begin{pmatrix} \cos\gamma & 0 & \sin\gamma \\ 0 & 1 & 0 \\ -\sin\gamma & 0 & \cos\gamma \end{pmatrix} \cdot \begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} \tag{9.20}$$

Rotation of β about the X"-axis

The final stage is to consider the rotation β about the X"-axis. From Fig. 9.07 the rotation matrix is

$$\mathbf{R}_X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\beta & \sin\beta \\ 0 & -\sin\beta & \cos\beta \end{pmatrix} \tag{9.21}$$

so that

$$\begin{pmatrix} X^* \\ Y^* \\ Z^* \end{pmatrix} = \mathbf{R}_X \cdot \begin{pmatrix} X'' \\ Y'' \\ Z'' \end{pmatrix} \tag{9.22}$$

$$= \mathbf{R}_Z \mathbf{R}_Y \cdot \mathbf{R}_X \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{9.23}$$

In the general case of defining the aspect of a map projection there may be rotation about one, two or all three axes.

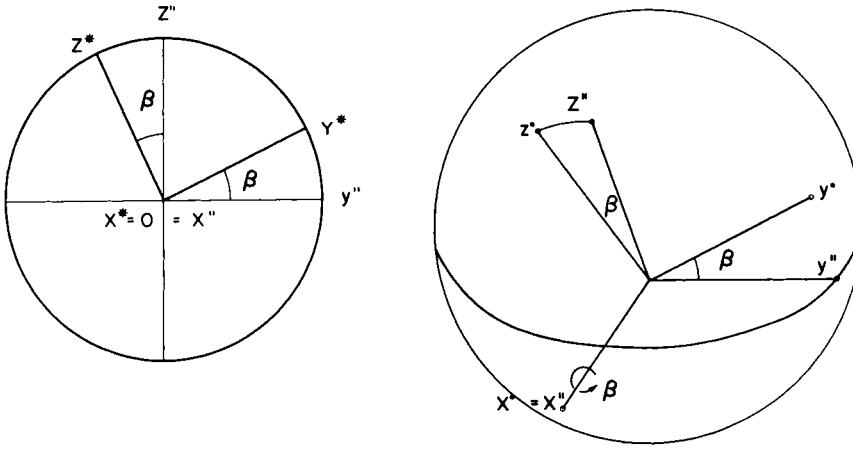


FIG. 9.08 Definition of the Eulerian angle β.

Thus the combination of all the rotations, in the order listed is

$$\mathbf{R} = \mathbf{R}_Z \,.\, \mathbf{R}_Y \,.\, \mathbf{R}_X = \begin{pmatrix} \cos\alpha \,.\, \cos\beta \,.\, \cos\gamma - \sin\alpha \,.\, \cos\gamma & -\cos\alpha \,.\, \cos\beta \,.\, \sin\gamma - \sin\alpha \,.\, \cos\gamma & \cos\alpha \,.\, \sin\beta \\ \sin\alpha \,.\, \cos\beta \,.\, \cos\gamma + \cos\alpha \,.\, \cos\gamma & -\sin\alpha \,.\, \cos\beta \,.\, \sin\gamma + \cos\alpha \,.\, \cos\gamma & \sin\alpha \,.\, \sin\beta \\ -\sin\beta \,.\, \cos\gamma & \sin\beta \,.\, \cos\gamma & \cos\beta \end{pmatrix} \quad (9.24)$$

and so

$$\begin{pmatrix} X_{\alpha\beta\gamma} \\ Y_{\alpha\beta\gamma} \\ Z_{\alpha\beta\gamma} \end{pmatrix} = \mathbf{R} \,.\, \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \quad (9.25)$$

The formulae that represent the nine elements of **R** are applicable for all values of α, β, γ between $+180°$ and $-180°$; they relate to a specific sequence of rotations, and will be different for another sequence. This means that the rotations applied to determine the aspect of a map projection ought to be carried out in a definite order; that preferred uses the alphabetical sequence, α, β, γ identified by the three Eulerian angles as employed by Wray (1974). Thus the rotation about the Z-axis is the primary rotation and that about the Y-axis is the secondary rotation, so that the result in equation (9.21) is the form which is required. However, the full matrix is not required in that form.

Wray's use of the Eulerian angles

Although Wray starts with these three Eulerian angles he changes the notation to be more appropriate definitions for his seven different aspects. Thus he introduces the three angles Φ, Λ, Ω, which are illustrated in Fig. 9.09. From this diagram it can be seen that the relationship between the angles α, β, γ and Λ, Φ, Ω which he calls the aspect parameters are:

$$\alpha \ldots \Lambda \pm 180^\circ$$
$$\beta \ldots 90^\circ - \Phi$$
$$\gamma \ldots -\Omega$$

It follows, moreover, that the full rotation matrix is only required for oblique skew and plagal aspects, corresponding to Fig. 7.04(d), (e) and (g). In all other aspects one or other of the aspect parameters are equal to 0°, 90° or 180°, so that the corresponding trigonometric functions are either equal to zero or unity. If they are zero, the rotation matrix is soon reduced to only a few simple terms. Table 9.01 records the values of the aspect parameters for the seven examples illustrated in Fig. 7.04. The numerical values given in the three right-hand columns refer to the examples illustrated in that figure.

It is also instructive to apply these rules to some of the examples listed in the book. For example, Briesemeister's projection, which was described on p. 156, has the definition $\Lambda = -165^\circ$, $\Phi = +45^\circ$, $\Omega = 0$.

Having established the form of the aspect parameters, it is necessary

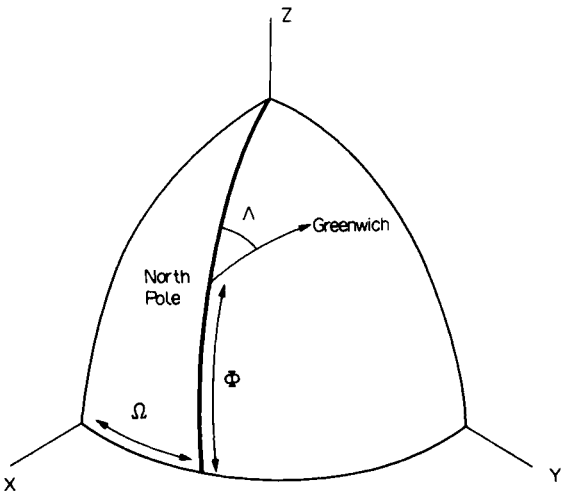


FIG. 9.09 The fundamental trirectangular spherical triangle employed by Wray to define the seven aspects of a map projection. (Source: Wray, 1974.)

TABLE 9.01 *The seven aspects of map projections: values of the aspect parameters*

Aspect	Major axis	Minor axis	Top	Bottom	Λ	Φ	Ω
Normal	Equator	Central Meridian	N. Pole	S. Pole	–	+90°	–
Simple Oblique	–	Central Meridian	–	–	–	+30°	0°
First Transverse	Central Meridian	–	Equator	Equator	–	0°	0°
Transverse Oblique	Central Meridian	–	Equator	Equator	−120°	0°	45°
Second Transverse	Central Meridian	Equator	Equator	Equator	+160°	0°	−90°
Equiskew	–	–	–	–	+90°	+45°	−90°

to use the results to obtain a new projection. One way of doing this is to use the dynamics concept of a *rigid-body rotation* of the vector OA through the three Eulerian angles.

First we must convert the geographical coordinates of graticule intersections from their geographical coordinates into three-dimensional cartesian coordinates. This is done by the three following equations:

$$X = \cos\varphi \,.\, \sin\lambda \tag{9.26}$$

$$Y = \cos\varphi \,.\, \cos\lambda \tag{9.27}$$

$$Z = \sin\varphi \tag{9.28}$$

Secondly, we apply the appropriate rotations according to equation (9.24) we convert them into the (X*, Y*, Z*) system which has been defined above. It is then necessary to convert back into geographical coordinates and proceed with the use of the map projection equations as if we were computing the master grid coordinates for the normal aspect.

Solution by the Cayley–Rodrigues method

The orthogonal matrix may also be formed from another set of independent parameters, which are associated with the direction cosines of the *one* axis of rotation needed to effect the transformation. This is useful and economical in digital processing because trigonometrical functions are not required to establish the matrix in the production of equations such as (9.15)–(9.24). The matrix is known as *Rodrigues matrix*, after the mathematician who devised it in 1840, and it was known to Cayley (1843) when he investigated the mapping of a sphere upon itself in one of the earliest of all his papers. We therefore describe this as the *Cayley–Rodrigues* method. It had not been used in cartography until it was employed by Barton (1976) and Arthur (1978).

Consider the trirectangular spherical triangle XYZ which is subjected to the rotations which we have already considered in detail. After the third rotation the vertices have moved to the new positions, X_0'', Y_0'', Z_0''. Because this triangle still represents the positions of the three coordinate axes, the shape and size of it is unaltered; but the position of it upon the spherical surface has been shifted, and each of the vertices may have moved a different amount. In this process of shifting, each of the vertices traces the arc of a great circle, X_0X_0', Y_0Y_0' and Z_0Z_0'. If, now, we bisect each of these arcs and construct the perpendicular arcs to them, we find that they define a single pole, *P*.

At this pole *P*, the spherical angle formed by the pair of secondaries defining the shift in each vector is the angle ξ. Thus the spherical angles:

$$X_0PX_0' = \xi \tag{9.29}$$

$$Y_0PY_0' = \xi \tag{9.30}$$

$$Z_0PZ_0' = \xi \tag{9.31}$$

The next property of interest to us is that the arcs between P and the ends of each of the great circle elements are equal. Thus

$$X_0P = X_0'P = f \tag{9.32}$$

$$Y_0P = Y_0'P = g \tag{9.33}$$

$$Z_0P = Z_0'P = h \tag{9.34}$$

and these are the direction cosines of the axis PP' with respect to the (X, Y, Z) or (X′, Y′, Z′) axes. Therefore

$$\cos^2 f + \cos^2 g + \cos^2 h = 1 \tag{9.35}$$

We may then write

$$p = \tan \tfrac{1}{2}\xi \,.\cos f \tag{9.36}$$

$$q = \tan \tfrac{1}{2}\xi \,.\cos g \tag{9.37}$$

$$r = \tan \tfrac{1}{2}\xi \,.\cos h \tag{9.38}$$

In analytical geometry the direction cosines are usually labelled *l*, *m*, *n*. We assign independent parameters λ, μ, ν corresponding to these and form the matrix:

$$\mathbf{R} = 1/\Delta \begin{pmatrix} 1+\frac{1}{4}(\lambda^2-\mu^2-\nu^2) & \nu+\frac{1}{2}\mu\lambda & -\mu+\frac{1}{2}\nu\lambda \\ -\nu+\frac{1}{2}\lambda\mu & \mu+\frac{1}{2}\lambda\nu & 1+\frac{1}{4}(-\lambda^2+\mu^2-\nu^2) \\ -\lambda+\frac{1}{2}\mu\nu & \lambda+\frac{1}{2}\nu\mu & 1+\frac{1}{4}(-\lambda^2-\mu^2+\nu^2) \end{pmatrix} \tag{9.39}$$

The Rodrigues Matrix is usually written in this form. See Thompson (1969) for an account of its role in matrix algebra. Cayley introduced the

terms in p, q and r in the same order as λ, μ, ν appear in (9.39), only omitting the convention of the parentheses enclosing the matrix. Substitution of these terms and multiplying through by $1/\Delta$, where:

$$\Delta = 1+p^2+q^2+r^2 \tag{9.40}$$

We have

$$X' = (1+p^2-q^2-r^2)X/\Delta+2(r+pq)Y/\Delta+2(-q+pr)Z/\Delta \tag{9.41}$$

$$Y' = 2(-r+qp)X/\Delta+(1-p^2-q^2-r^2)Y/\Delta+2(p+qr)Z/\Delta \tag{9.42}$$

$$Z' = 2(q+rp)X/\Delta+2(-p+rq)Y/\Delta+(1-p^2-q^2+r^2)Z/\Delta \tag{9.43}$$

The amount of arithmetic involved is appreciably less than in the method using trigonometrical functions. The subroutines used in digital computers are of the form of equation (4.26), p. 73, so that the absence of series expansions reduces the amount of arithmetic involved for the powers and multiples of p, q and r only have to be determined once for a particular aspect. Each of the equations (9.37)–(9.40) involves hardly any more arithmetic than finding the sine or the cosine of a single angle by other methods. It should be noted that the majority of the angles used in the Cayley–Rodrigues solution are in the third or fourth quadrant, because the point P lies on the other side of the sphere and most of them must be obtained by subtracting latitude or longitude from 360°. Thus, for the simple oblique projection with origin $\varphi_0 = 60°$, for which, in Table 9.01, Wray assigns $\Phi = 30°$, $\Omega = 0°$, we have p = 0·0, q = −0·55735, r = 0·0, $\theta = 300°$. The resulting rotation matrix is:

$$\mathbf{R} = \begin{pmatrix} 0{\cdot}5 & 0 & 0{\cdot}866 \\ 0 & 1 & 0 \\ -0{\cdot}866 & 0 & 0{\cdot}5 \end{pmatrix} \tag{9.44}$$

For the first transverse aspect where $\Lambda = 120°$, $\Phi = 0°$, $\Omega = 45°$, we have p = 0·0, q = −1, r = 0·0, and $\theta = 270°$. The resulting matrix is

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix} \tag{9.45}$$

These examples indicate that the rotation matrix usually simplifies to an easily remembered combination of elements. Moreover, it is constant for any projection, varying only with aspect. Thus the matrix

$$\mathbf{R} = \begin{pmatrix} \sqrt{2}/2 & 0 & \sqrt{2}/2 \\ 0 & 1 & 0 \\ -\sqrt{2}/2 & 0 & \sqrt{2}/2 \end{pmatrix} \tag{9.46}$$

describes the rotation of the axes for *all* simple oblique aspect projections having $\varphi_0 = 45°$.

As an example, we take the same point in Briesemeister's projection which we used earlier; namely $\varphi = 30°\text{N}$, $\lambda = 75°\text{E}$. Since the origin of the projection is $\varphi_0 = 45°\text{N}$, $\lambda_0 = 15°\text{E}$, the difference in longitude $\delta\lambda = 60°$.

This point may be converted into the three-dimensional cartesian coordinates, using equations (9.22), (9.23) and (9.24). Thus $X = 0{\cdot}4330127$, $Y = 0{\cdot}75$ and $Z = 0{\cdot}5$. Applying the rotations of (9.43) we find the rotated axes are:

$$X^* = 0{\cdot}65974,\ Y^* = 0{\cdot}75,\ Z^* = 0{\cdot}04737.$$

Converting these back into spherical polar coordinates, we find:

$$\varphi' = 2{\cdot}7149,\ \lambda' = 48{\cdot}6634$$

Finally we turn to the projection equations given on p. 440 and find that

$$x = 0{\cdot}7878,\ y = 0{\cdot}0518.$$

These values correspond to the Briesemeister projection coordinates for that point in Appendix II.

CHAPTER 10

The analytical derivation of some map projections

> Geography is not alone in the embarrassing abundance of its material; the mammalia are only one, and not the largest, of sixteen classes of animals and there are about 5,000 species of mammals alone; merely to read a list of their names would waste about three lecture hours, yet with this vast unexplored field of mammalian zoology awaiting investigation the zoology student spends about sixty hours dissecting the rabbit – and with profit. There is something here for us to ponder. Should we not be gaining more valuable discipline if we took much of the routine description for granted and employed our time in dissecting the anatomy of a map as thoroughly and exhaustively as he dissects a rabbit, and like him, in getting down to the guts of the matter?
>
> A. A. Miller, *Presidential Address to the Institute of British Geographers*, 1948

Introduction

Throughout this book we have been preoccupied with principles and with practical techniques. Although we have referred to, and illustrated, a variety of different map projections, we have not yet attempted to derive any of the coordinate expressions needed to define and construct a particular projection. An exception might be made of the Cylindrical equal-area projection which has been described in some detail in Chapters 5, 6 and 7. But even with this example it was taken on trust that the Cylindrical equal-area projection satisfied the special property described in its title, at least up to the stage of tabulating the particular scales and distortion characteristics in Table 6.01, p. 111. Up to that stage the reader just had to accept that this was so, simply because we had made this assertion. This is a fundamental weakness in most descriptive studies of map projections where the method of presentation is primarily geometrical. Almost as an afterthought we are told that a particular projection is conformal or equal-area, or more commonly that it has 'arbitrary properties'. To the beginner this means that names and properties have to be correctly equated and committed to memory, for many of the most popular projections provide no clue in their names to any special property or where

they belong in a classification system. It is therefore necessary to memorise the facts that the *Stereographic* and *Mercator's* projections are conformal; that *Bonne's* or *Mollweide's* are equal-area; that the first is azimuthal, the second is cylindrical, the third is pseudoconical or the fourth is pseudocylindrical. To the intelligent beginner, the scientist or the engineer, it may seem that the subject of map projections is an empirical ragbag of unrelated facts which appear to have been collected almost accidentally, and that there is no particular thread of continuity in the processes of reasoning through which they have been derived.

We therefore believe that it is both desirable and necessary to demonstrate the *analytical approach* to the study of map projections. In other words, we must show how it is possible to derive a map projection which satisfies a particular property within the limitations imposed by the group and class to which it belongs. Thus we start by stating certain initial mathematical constraints and finish with the coordinate equations for the map projection which satisfies them; with a table of the distortion characteristics as well.

We do not intend to give a comprehensive analysis of all the special properties which can be derived in every class of projection. We may learn much about the analytical approach from the study of a few well-known examples from Group D of the classification system.

Example I: The azimuthal equal-area projection (Lambert)

The first example illustrates how an azimuthal projection may be derived which satisfies the special properties of equivalence.

We have already seen in Chapter 7 that the azimuthal class is one of the subdivisions of Group D. Moreover it has been specified that the azimuthal projections can be defined in terms of polar coordinates according to the special condition that the origin of these coordinates is the only point of zero distortion at the centre of the map. In the normal aspect this point is the geographical pole. These conditions have the geometrical significance of being the transformation of the spherical surface to a plane which is tangential to it at the geographical pole, as shown in Fig. 5.07, p. 90.

Conditions applicable to any azimuthal projection

It follows from the definition of a spherical angle in Chapter 3, p. 54, as well as from the creation of a point of zero distortion at the origin of polar coordinates, that any plane angle θ at that point is equal to the corresponding angle on the globe. At the pole this spherical angle represents longitude; therefore we may write the first of the essential equa-

tions to define *any azimuthal projection* in its normal or polar aspect as

$$\theta = \lambda \tag{10.01}$$

From the functional relationships which have been established in Chapter 5 we also know that in a normal aspect azimuthal projection the meridians are represented by straight lines and the parallels are concentric circles with common centre at the pole. Since the parallels of latitude must satisfy the functional relationship of Group D that $\mathrm{r} = f_1(\varphi)$, and since θ has already been determined, it follows that the only way in which we may derive an azimuthal projection to satisfy a special property is to seek a suitable expression for the radius of each parallel. We may also write

$$\mathrm{r} = f(\chi) \tag{10.02}$$

where χ is the colatitude. We have made this change because it is algebraically simpler to derive an expression for the radius vector in terms of colatitude. Moreover this facilitates conversion from the equations derived for the normal aspect into the general expressions needed to construct the projection in any aspect using bearing and distance coordinates.

Since the graticule intersections of the normal aspect are orthogonal, it follows that the principal directions coincide with the graticule and therefore the particular scales along the meridian and parallel are the maximum and minimum particular scales at a point. Consequently we have the alternative conditions that, either

$$h = a \text{ and } k = b \tag{10.03}$$

or

$$k = a \text{ and } h = b \tag{10.04}$$

The analysis of the particular scales can be made from comparison of the infinitely small corresponding figures $ABCD$ on the spherical surface and $A'B'C'D'$ on the plane. In Chapter 5 we developed these arguments for the general case of any map projection. Now we modify them according to the special conditions common to any azimuthal projection. Figure 10.01(a) illustrates the portion of the spherical surface in which the two parallels are φ and $\varphi + \mathrm{d}\varphi$, and the meridians are λ and $\lambda + \mathrm{d}\lambda$. The radii of the parallels on the map, Fig. 10.01(b), are $N'A = \mathrm{r}$ and $N'B = \mathrm{r} - \mathrm{dr}$. The vectorial angle $A'N'D' = \mathrm{d}\theta$.

The scale along the meridian through A' is the relationship $A'B'/AB$ or

$$h = -\mathrm{dr}/\mathrm{R} \,.\, \mathrm{d}\varphi \tag{10.05}$$

$$= \mathrm{dr}/\mathrm{R} \,.\, \mathrm{d}\chi \tag{10.06}$$

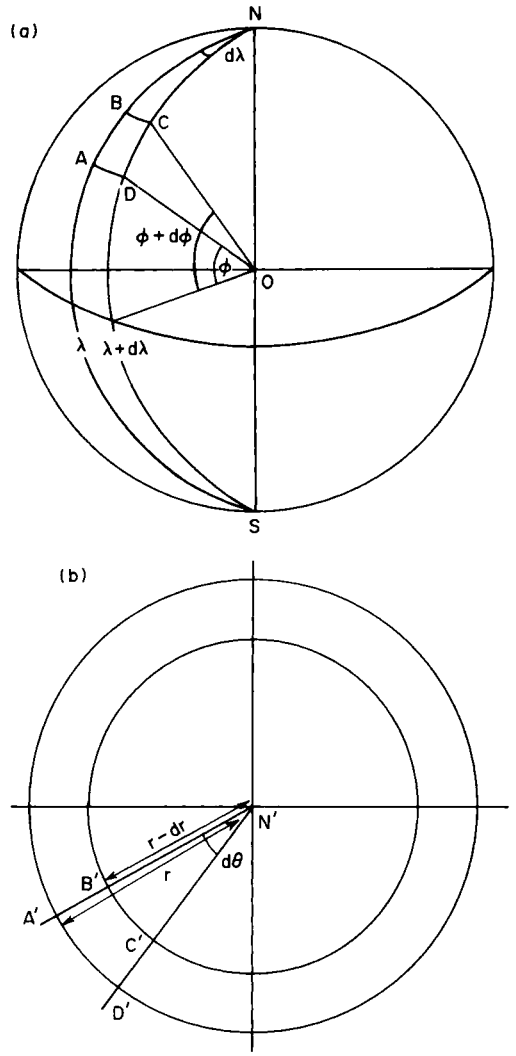


FIG. 10.01 An infinitely small quadrilateral, $ABCD$, on the spherical surface and its plane representation $A'B'C'D'$ by means of an Azimuthal projection.

Note that if we use the expression for latitude we must allocate the negative sign to dr, because r increases as latitude decreases. In equation (10.06) dr is positive because r increases with colatitude.

The scale along the parallel through A' is the relationship $A'D'/AD$ or

$$k = \mathrm{r} \,.\, \mathrm{d}\theta/\mathrm{R} \,.\, \cos\varphi \,.\, \mathrm{d}\lambda \qquad (10.07)$$

$$= \mathrm{r} \,.\, \mathrm{d}\theta/\mathrm{R} \,.\, \sin\chi \,.\, \mathrm{d}\lambda \qquad (10.08)$$

Since we have already specified that $\theta = \lambda$, we may also put $d\theta = d\lambda$ so that (10.08) simplifies to

$$k = r/R \,.\, \sin\chi \tag{10.09}$$

We have already noted that the principal directions coincide with the graticule. Therefore

$$\begin{aligned} p &= h \,.\, k \\ &= [dr/R \,.\, d\chi] \,.\, [r/R \,.\, \sin\chi] \end{aligned} \tag{10.10}$$

and

$$\sin(\omega/2) = |h-k|/(h+k) \tag{10.11}$$

to give the equation for maximum angular deformation. In equation (10.11) we use the modulus $|h-k|$ to denote the positive difference between the larger and smaller values for h and k which are, as yet, unspecified. This is the same as writing $h \sim k$.

The special conditions for the azimuthal equal-area projection

Equations (10.01)–(10.11) apply equally to all normal aspect azimuthal projections. We wish to obtain an equal-area projection. From equation (6.27) this is the condition that $a \,.\, b = 1$. However we have seen that in the normal aspect azimuthal projections, $h \,.\, k = a \,.\, b$; therefore we can satisfy the property of equivalence by making the right-hand side of equation (10.10) equal to unity, i.e.

$$[dr/R \,.\, d\chi] \,.\, [r \,.\, R \,.\, \sin\chi] = 1 \tag{10.12}$$

or

$$r \,.\, dr = R^2 \,.\, \sin\chi \,.\, d\chi \tag{10.13}$$

This must be solved by integration of r with respect to χ, i.e.

$$\tfrac{1}{2}r^2 = R^2 \int_0^{\chi} \sin\chi \,.\, d\chi \tag{10.14}$$

From elementary calculus, the integral $\int \sin\theta \,.\, d\theta = -\cos\theta + C$, where C is the *integration constant*. Therefore

$$\begin{aligned} r^2 &= -2R^2 \,.\, \cos\chi + C \\ &= C - 2R^2 \,.\, \cos\chi \end{aligned} \tag{10.15}$$

In the normal aspect azimuthal equal-area projection, where $\chi = 0$, $r = 0$, $\cos\chi = 1{\cdot}0$ so that $C - 2R^2 = 0$ and $C = 2R^2$. Consequently

$$r^2 = 2 \,.\, R^2(1 - \cos\chi) \tag{10.16}$$

There is an algebraic manipulation in trigonometry that

$$1 - \cos\theta = 2\sin^2(\theta/2)$$

Therefore (10.16) may be expressed in the form

$$r^2 = 4 \,.\, R^2 \sin^2(\chi/2) \tag{10.17}$$

so that, finally,

$$r = 2 \,.\, R \,.\, \sin(\chi/2) \tag{10.18}$$

Equations (10.01) and (10.18) are the two equations needed to define the normal aspect azimuthal equal-area projection in polar coordinates. For a spherical earth of unit radius, these may be written in the form

$$\begin{aligned} r &= 2 \,.\, \sin(\chi/2) \\ \theta &= \lambda \end{aligned} \tag{10.19}$$

Once a value for r has been determined, this may be substituted in the general expressions for the particular scales. Thus, replacing r in equations (10.06) and (10.09) by the right-hand side of (10.18) we obtain

$$h = \cos\chi/2 \tag{10.20}$$

$$k = \sec\chi/2 \tag{10.21}$$

TABLE 10.01 *Normal aspect Azimuthal equal-area projection (Lambert). Table of radii of parallels, particular scales and distortion characteristics of 15° increments in latitude*

Latitude φ	Radius r	Particular scales h	Particular scales k	Area scale p	Maximum angular deformation $\omega°$
0°	1·4142	0·7071	1·4142	1·0000	38°57′
15°	1·2175	0·7934	1·2605	1·0000	26°17′
30°	1·0000	0·8660	1·1547	1·0000	16°26′
45°	0·7654	0·9239	1·0824	1·0000	9°04′
60°	0·5176	0·8659	1·0353	1·0000	3°58′
75°	0·2611	0·9914	1·0086	1·0000	0°59′
90°	0·0000	1·0000	1·0000	1·0000	0°

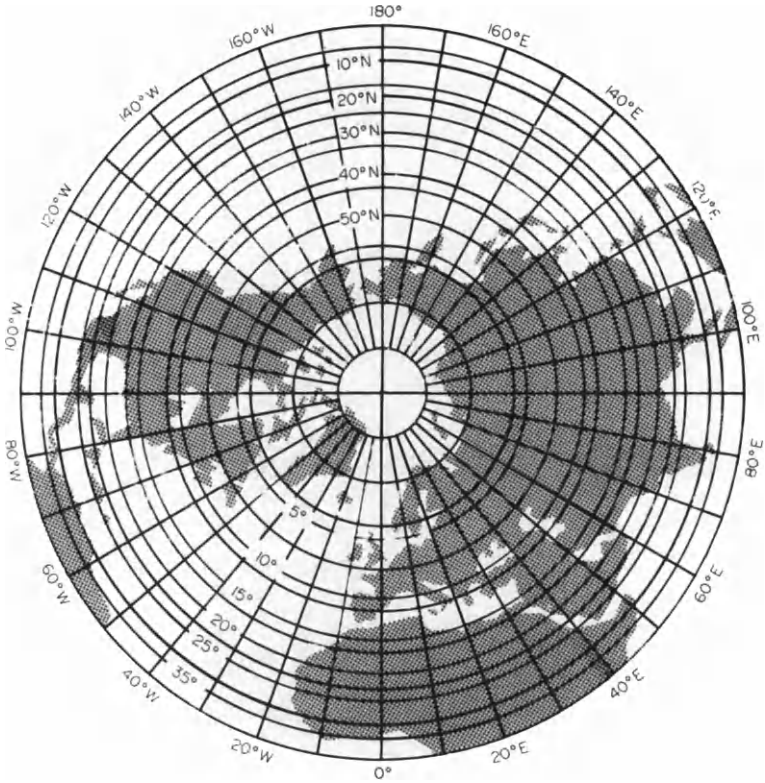


FIG. 10.02 The normal aspect of the Azimuthal equal-area projection (Lambert) (No. 12 in Appendix I). The origin of the projection is the North Pole. The isograms represent equal values of maximum angular deformation (ω) at 5°, 10°, 15°, 20°, 25° and 35°. These are identical to the isograms shown in Figs 10.03 and 10.04, except that the 30° isogram has been omitted for greater clarity.

Substituting these expressions in (10.11) we obtain numerical values for ω. Table 10.01 gives the results of these calculations. This table only shows numerical values for a hemispherical map, but the projection may be extended to show the entire world. In this case the boundary represents the antipodal point of the origin (the South Pole in the normal aspect with origin at the North Pole). This is a singular point which is mapped as the circumference of a circle of radius 2R.

As demonstrated in Chapter 8, the transverse and oblique aspects of the Azimuthal equal-area projection may be derived simply by substituting z for χ and α for λ in the foregoing equations. The three aspects of the projection are illustrated in Figs 10.02, 10.03 and 10.04.

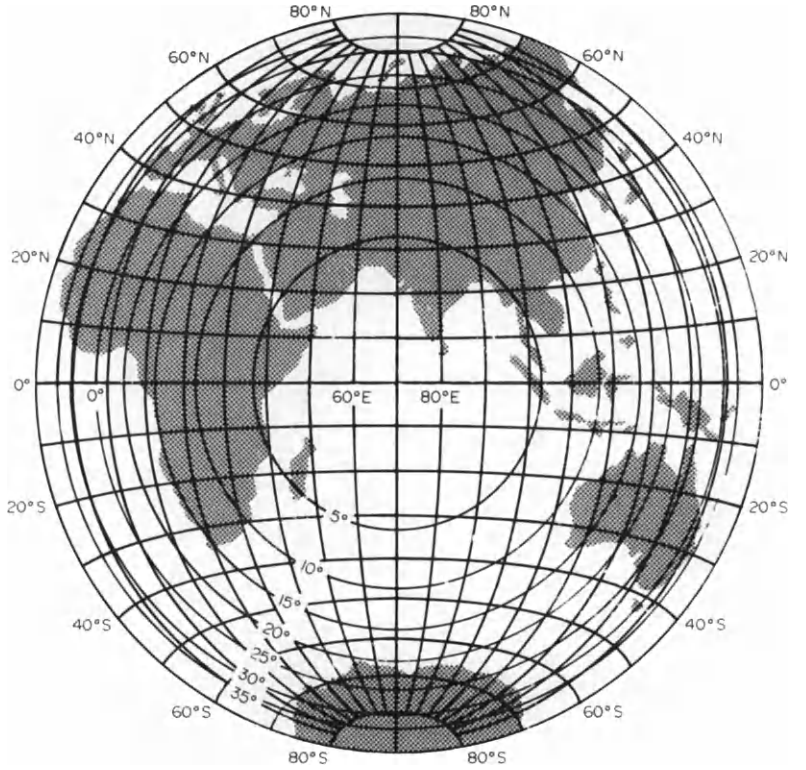


FIG. 10.03 The equatorial or transverse aspect of the Azimuthal equal-area projection. The origin of the projection is on the equator in longitude 70°E. The isograms represent equal values of maximum angular deformation (ω) at 5° intervals. These are identical to the isograms shown in Figs 10.02 and 10.04.

Example II: The conical equidistant projection with one standard parallel (Ptolemy) and the conical equidistant projection with two standard parallels (de l'Isle)

From Chapter 7 we know that the conical class of projections also belongs to Group D and, like the azimuthal projections, these may be derived in polar coordinates. The differences between these two classes are, first, that the origin of the polar coordinates used to define any conical projection in its normal aspect is not the geographical pole. Secondly, a fundamental property of all conical projections is that the line of zero distortion is one or two arcs of small circles. In the normal aspect this is one of two parallels of latitude, known as *standard parallels*.

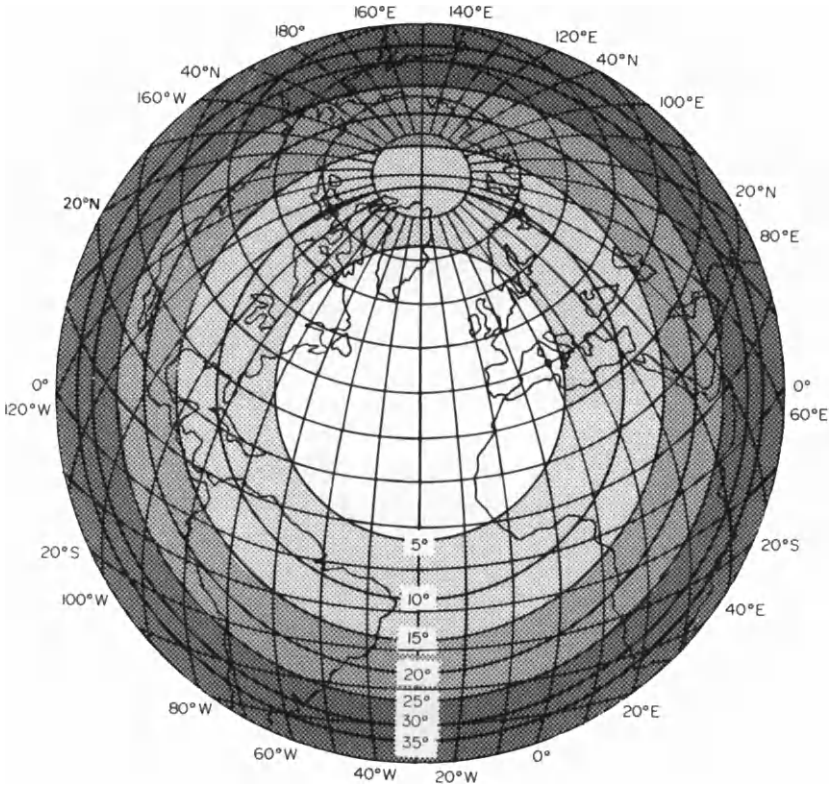


FIG. 10.04 The oblique aspect of the Azimuthal equal-area projection. The origin of the projection is in latitude 40°N, longitude 30°W. The isograms represent equal values of maximum angular deformation (ω) at 5° intervals. These are identical to the isograms shown in Figs 10.02 and 10.03.

Conditions applicable to all conical projections

From the brief description of the class in Chapter 7, we already know that the meridians of the normal aspect are represented by straight lines which converge to the origin of the polar coordinates. This point is usually located some distance beyond the geographical pole, as illustrated in Fig. 10.06 where it is represented by the vertex, V. This has the important effect of altering the relationship between the vectorial angle, θ, and longitude, so that there is a constant $n < 1$ of the form

$$\theta = n \, . \, \lambda \tag{10.22}$$

The term n is known as the *constant of the cone*.

The parallels of the normal aspect conical projection are concentric circular arcs having the common centre at the vertex. It follows that in many, though not all, conical projections, the geographical pole is represented by a short circular arc instead of a point. Such projections are sometimes described as *truncated conical projections* to distinguish them from those examples where the pole is a point. In all the truncated conical projections the pole is obviously a singular point. Both of the examples studied here belong to the truncated category.

Derivation of the particular scales for the conical projections follows arguments similar to those already employed on pp. 197–199. Figures 10.05(a) and (b) represent the slightly different meanings of r and θ. This time we will derive the equations in terms of latitude though, of course, this could be done through the argument of colatitude. We note that the conditions expressed by equations (10.03) and (10.04) still apply so that if we can derive the particular scales along the meridian and parallel through A', we have also obtained the maximum and minimum particular scales.

As in (10.05)

$$h = -\mathrm{dr}/\mathrm{R} \,.\, \mathrm{d}\varphi \tag{10.23}$$

and, as in (10.07)

$$k = \mathrm{r} \,.\, \mathrm{d}\theta/\mathrm{R} \,.\, \cos\varphi \tag{10.24}$$

However, following (10.22) we must now write $\mathrm{d}\theta = n \,.\, \mathrm{d}\lambda$ so that the expression for the particular scale along the parallel now becomes

$$k = n \,.\, \mathrm{r}/\mathrm{R} \,.\, \cos\varphi \tag{10.25}$$

It is now necessary to evaluate the constant of the cone. The first condition which defines it is that we have specified that the principal scale is preserved along the standard parallel. Thus, denoting the scale along the standard parallel by k_0 we must fulfil the condition that

$$k_0 = \mathrm{r}_0 \,.\, \mathrm{d}\theta/\mathrm{R} \,.\, \cos\varphi_0 \,.\, \mathrm{d}\lambda = 1 \tag{10.26}$$

where r_0 is the radius of the standard parallel in latitude φ_0. Therefore

$$\mathrm{d}\theta = [\mathrm{R} \,.\, \cos\varphi_0/\mathrm{r}_0] \,.\, \mathrm{d}\lambda \tag{10.27}$$

or

$$\theta = [\mathrm{R} \,.\, \cos\varphi_0/\mathrm{r}_0] \,.\, \lambda \tag{10.28}$$

The second condition which defines the constant of the cone is the geometrical requirement that the surface of a cone which is tangential to the spherical surface must be perpendicular to the radii along the small circle of contact. Hence in the triangle VAO illustrated in Fig. 10.06, the angle OAV is a right angle. Therefore

$$AM = \mathrm{r}_0 \,.\, \sin\varphi_0 \tag{10.29}$$

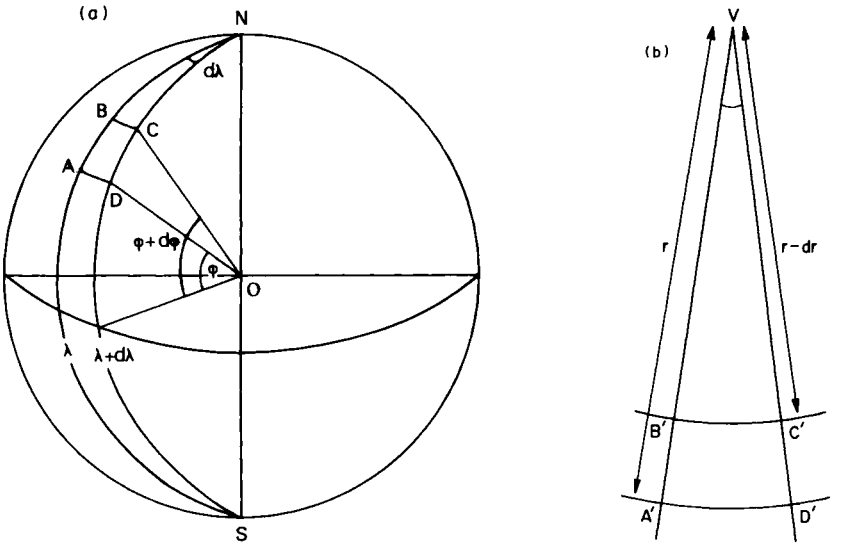


FIG. 10.05 An infinitely small quadrilateral, $ABCD$, on the spherical surface and its plane representation $A'B'C'D'$ by means of a conical projection.

But AM corresponds to the radius of the standard parallel which on the sphere is equal to $R \cdot \cos \varphi_0$. Substituting the right-hand side of (10.29) in (10.28)

$$\theta = [r_0 \cdot \sin \varphi_0 / r_0] \cdot \lambda \tag{10.30}$$

$$= \sin \varphi_0 \cdot \lambda \tag{10.31}$$

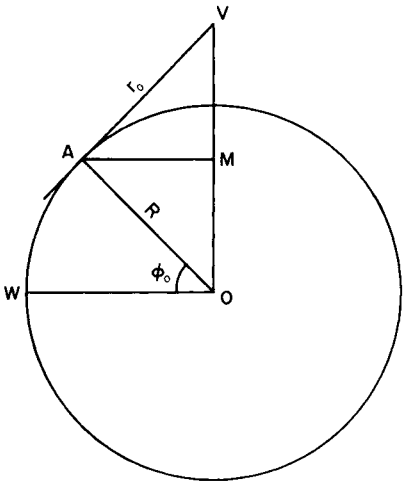


FIG. 10.06 The determination of the radius of the standard parallel, r_0, of a conical projection with one standard parallel.

Substituting this in equation (10.22)

$$n = \sin \varphi_0 \tag{10.32}$$

The special conditions of the equidistant conical projections

We wish to preserve the special property of equidistance, i.e. $h = 1$. Substituting this in (10.23)

$$h = -\mathrm{dr}/\mathrm{R} \,.\, \mathrm{d}\varphi = 1$$

and therefore

$$\mathrm{dr} = -\mathrm{R} \,.\, \mathrm{d}\varphi \tag{10.33}$$

Integration of this expression yields

$$\mathrm{r} = C - \mathrm{R} \,.\, \varphi \tag{10.34}$$

where C is the integration constant which may be interpreted as follows. In equation (10.34) we put $\varphi = 0$. Then $\mathrm{R} \,.\, \varphi = 0$ and therefore $\mathrm{r} = C$. In other words, this constant represents the radius of the equator on the projection. If we had proceeded, as in the study of the Azimuthal equal-area projection, to derive the Conical equidistant projection in terms of colatitude we would have obtained as the integration constant a value corresponding to the radius of the circular arc representing the geographical pole.

It now remains to relate the radius of any parallel to that of the standard parallel. This may be done analytically but is also easily found from Fig. 10.06, where it can be seen that the angle $AVO = MAO = AOW = \varphi_0$; therefore the radius, VA of the standard parallel is

$$\mathrm{r}_0 = \mathrm{R} \,.\, \cot \varphi_0 \tag{10.35}$$

We may now express the constant C in terms of the radius of the standard parallel. If C represents the radius of the equator

$$C = \mathrm{r}_0 + \mathrm{R} \,.\, \varphi_0 \tag{10.36}$$

$$= \mathrm{R} \,.\, \cot \varphi_0 + \mathrm{R} \,.\, \varphi_0 \tag{10.37}$$

From (10.34), therefore, the radius of any parallel may be written

$$\mathrm{r} = \mathrm{R} \,.\, \cot \varphi_0 + \mathrm{R}(\varphi_0 - \varphi) \tag{10.38}$$

This expression, together with (10.31) gives the polar coordinates for any point on this projection. For $\mathrm{R} = 1$, therefore, the equations defining the Conical equidistant projection (Ptolemy) are

$$\mathrm{r} = \cot \varphi_0 + (\varphi_0 - \varphi)$$

$$\theta = \sin \varphi_0 \,.\, \lambda \tag{10.39}$$

TABLE 10.02 *Conical equidistant projection with one standard parallel (Ptolemy). Numerical values for radii of parallels, particular scales and distortion characteristics of the 15° graticule with standard parallel* $\varphi_0 = 45°$

Latitude φ	Radius of parallel r	Particular scales h	Particular scales k	Area scale p	Maximum angular deformation ω
0°	1·7854	1·0000	1·2625	1·2625	13°19′
15°	1·5236	1·0000	1·1153	1·1153	6°15′
30°	1·2618	1·0000	1·0303	1·0303	1°42′
45°	1·0000	1·0000	1·0000	1·0000	0°
60°	0·7382	1·0000	1·0440	1·0440	2°28′
75°	0·4762	1·0000	1·3015	1·3015	15°03′
90°	0·2146	1·0000	∞	–	180°

The particular scales may be determined by substitution in the equation for r in the general expression for the particular scale along the parallel. Thus

$$h = 1$$

$$k = [\cos\varphi_0 + (\varphi_0 - \varphi) \,.\, \sin\varphi_0]/\cos\varphi \qquad (10.40)$$

Because this is an equidistant projection

$$p = k$$

and the maximum angular deformation may be evaluated from the equation

$$\sin(\omega/2) = (k-1)/(k+1) \qquad (10.41)$$

Table 10.02 gives numerical values of these parameters for the particular case of a projection with the standard parallel in latitude 45°. This projection is often called the *Simple conical projection with one standard parallel.* It is illustrated in Fig. 10.07.

The Conical equidistant projection with two standard parallels (de l'Isle)

An important modification to any of the conical projections is to replace the single standard parallel with two. This is equivalent to the geometrical concept of the secant cone illustrated in Fig. 5.09, p. 92, which has the effect of redistributing the particular scales because the principal scale is now preserved along two parallels of latitude. This means that a greater extent in latitude may be mapped without excessive distortion.

In order to demonstrate this important principle to the projection

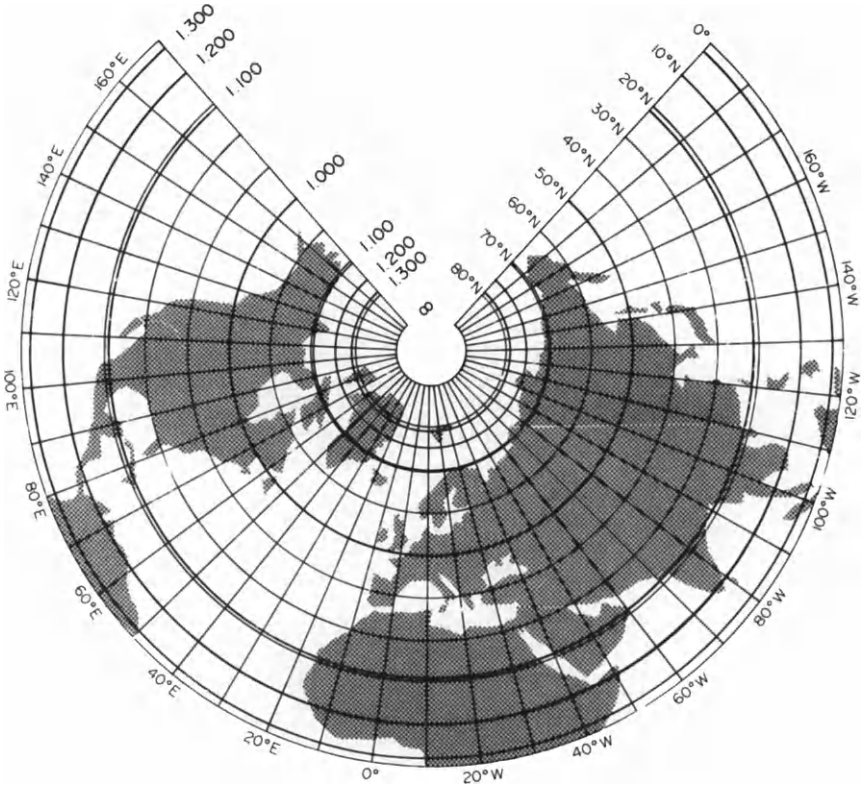


FIG. 10.07 The normal aspect of the Conical equidistant projection with one standard parallel (Ptolemy). In this map of the northern hemisphere the standard parallel is latitude 50°N. The isograms represent equal values of particular scale along the parallels, and since the particular scale along the meridians is everywhere equal to unity the numerical values for the isograms also represent area scale (p).

which has already been described, we explain the derivation of the *Conical equidistant projection with two standard parallels (de l'Isle)*, which is also known as the *Simple conical projection with two standard parallels.* Since Maling (1960) has shown that there are other equidistant conical projections, it is necessary to state for the de l'Isle projection that *the two standard parallels are located in latitudes which lie midway between the centre of the map and its bounding parallels.* Thus, if we desired to prepare a map of the northern hemisphere with bounding parallels $\varphi_S = 0°$ and $\varphi_N = 90°$, the central parallel, φ_0 is latitude 45°N and the two standard parallels lie in latitudes $\varphi_1 = 67°30'$N and $\varphi_2 = 22°30'$N.

Algebraically we may express these conditions as follows:

$$\varphi_1 = \varphi_N - 1/4(\varphi_N - \varphi_S) \tag{10.42}$$

$$\varphi_2 = \varphi_S + 1/4(\varphi_N - \varphi_S) \tag{10.43}$$

Since φ_1 and φ_2 are standard parallels, the particular scales along them are equal to unity. Thus

$$k_1 = k_2 = 1{\cdot}0 \tag{10.44}$$

We may obtain an equation containing k and the two constants n and C by combining equations (10.25) and (10.34):

$$k = [n(C - \mathrm{R} \,.\, \varphi)]/[\mathrm{R} \,.\, \cos\varphi] \tag{10.45}$$

For the two standard parallels this may be written as:

$$[n(C - \mathrm{R} \,.\, \varphi_1)]/[\mathrm{r} \,.\, \cos\varphi_1] = [\mathrm{n}(C - \mathrm{R} \,.\, \varphi_2)]/[\mathrm{R} \,.\, \cos\varphi_2] = 1 \tag{10.46}$$

This gives us the two solutions

$$C = \mathrm{R} \,.\, \varphi_1 + [\mathrm{R} \,.\, \cos\varphi_1/n] \tag{10.47}$$

$$C = \mathrm{R} \,.\, \varphi_2 + [\mathrm{R} \,.\, \cos\varphi_2/n] \tag{10.48}$$

Subtracting (10.48) from (10.47) and putting R = 1,

$$(\varphi_1 - \varphi_2) = [\cos\varphi_2 - \cos\varphi_1]/[\varphi_1 - \varphi_2] \tag{10.49}$$

or

$$n = [\cos\varphi_2 - \cos\varphi_1]/[\varphi_1 - \varphi_2] \tag{10.50}$$

From (10.46) it can also be shown that

$$[C - \mathrm{R} \,.\, \varphi_1]/[\mathrm{R} \,.\, \cos\varphi_1] = [C - \mathrm{R} \,.\, \varphi_2]/[\mathrm{R} \,.\, \cos\varphi_2] \tag{10.51}$$

which may be solved for C as

$$C = [\varphi_1 \,.\, \cos\varphi_2 - \varphi_2 \,.\, \cos\varphi_1]/[\cos\varphi_2 - \cos\varphi_1] \tag{10.52}$$

These new values for n and C may be used with equations (10.22) and (10.34) to construct the de l'Isle projection and determine its distortion characteristics. Numerical values for these are given in Table 10.03 and the projection is illustrated in Fig. 10.08.

Example III: Cylindrical Conformal, or Mercator's projection

We now examine the derivation of one of the most important of all map projections which, in the normal aspect, is the basis of most nautical charts and in the transverse aspect is equally important for topographical mapping. We defer examination of how the projection is used for these purposes until Chapters 14 and 16. Here we confine our attention to the derivation of it as the conformal member of the cylindrical class of map projections.

TABLE 10.03 *Conical equidistant projection (de l'Isle) with two standard parallels. Numerical values for radii of parallels, particular scales and distortion characteristics for the 15° graticule with standard parallels* $\varphi_1 = 67°30'N$ *and* $\varphi_2 = 22°30'N$; n = *0·68907;* C = *1.73346*

Latitude φ	Radius of parallel r	Particular scales h	Particular scales k	Area scale p	Maximum angular deformation (ω)
0°	1·7335	1·0000	1·1945	1·1945	10°10′
15°	1·4717	1·0000	1·4717	1·4717	2°47′
22°30′	1·3408	1·0000	1·0000	1·0000	0°
30°	1·2099	1·0000	0·9627	0·9627	2°11′
45°	0·9481	1·0000	0·9239	0·9239	4°32′
60°	0·6863	1·0000	0·9458	0·9458	3°12′
67°30′	0·5554	1·0000	1·0000	1·0000	0°
75°	0·4245	1·0000	1·1301	1·1301	7°0′
90°	0·1627	1·0000	∞	–	180°

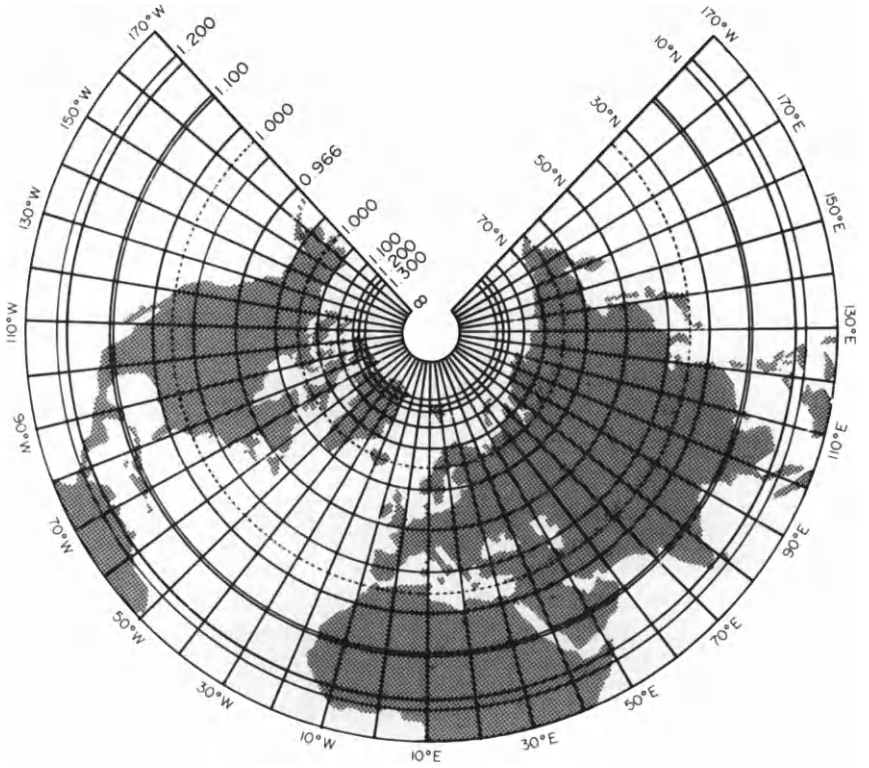


FIG. 10.08 The normal aspect of the Conical equidistant projection with two standard parallels. In this map of the northern hemisphere the standard parallels occur in latitudes 35°N and 65°N. The isograms represent equal values of particular scale along the parallels, and since the particular scales along the meridians are everywhere equal to unity the numerical values for the isograms also represent area scale (p).

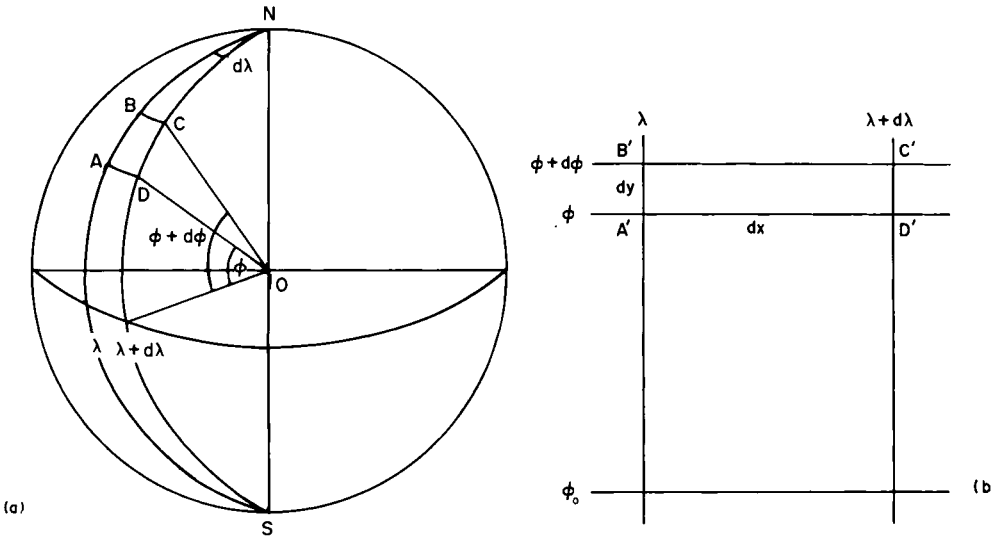


FIG. 10.09 An infinitely small quadrilateral, $ABCD$, on the spherical surface and its plane representation $A'B'C'D'$ by means of a cylindrical projection.

Conditions applicable to all cylindrical map projections

Figure 10.09(a) illustrates the representation of an infinitely small quadrangle on the spherical surface, and Fig. 10.09(b) illustrates the corresponding figure on the plane. Using the kind of argument to which the reader should now be accustomed, we may define the particular scale along the meridian at A' as

$$h = \mathrm{dy}/\mathrm{R} \,.\, \mathrm{d}\varphi \tag{10.53}$$

and the particular scale along the parallel as

$$k = \mathrm{dx}/(\mathrm{R} \,.\, \cos\varphi \,.\, \mathrm{d}\lambda) \tag{10.54}$$

The values of dx depend upon the spacing of the meridians on the map. Since the normal aspect of a cylindrical projection has only one line of zero distortion at the equator, this means that the meridians must be correctly spaced along the equator. In other words the equation

$$\mathrm{x} = \mathrm{R} \,.\, \lambda \tag{10.55}$$

is true for all normal aspect cylindrical projections which have not been modified. Substituting the corresponding expression for infinitely small increments in longitude in equation (10.54)

$$k = \mathrm{R} \,.\, \mathrm{d}\lambda/[\mathrm{R} \,.\, \cos\varphi \,.\, \mathrm{d}\lambda] \tag{10.56}$$

which simplifies to

$$k = 1/\cos\varphi \tag{10.57}$$

$$= \sec\varphi \tag{10.58}$$

In other words, the scale along the parallel varies according to the secant of the latitude. This, too, is *common to all normal aspect cylindrical projections.*

It follows from the pattern of parallels and meridians of the normal aspect cylindrical projection, which we remember is composed of two families of straight lines intersecting orthogonally, that the conditions described by equations (10.03) and (10.04) remain valid.

Conditions applicable to the Cylindrical Conformal projection

Thus we may simplify the algebraic condition for conformality, that $a = b$, with the expression

$$h = k \tag{10.59}$$

In other words, we put

$$\mathrm{dy/R}\,.\,\mathrm{d}\varphi = \mathrm{dx/R}\,.\cos\varphi \tag{10.60}$$

and solve this equation for y.

Equation (10.60) may be written in the form

$$\mathrm{dy/dx} = \mathrm{d}\theta/\cos\varphi \tag{10.61}$$

so that

$$\mathrm{y} = \int \sec\varphi\,.\,\mathrm{d}\varphi \tag{10.62}$$

The solution of this integral is well known in elementary calculus. Therefore we may write

$$\mathrm{y} = \ln\tan(\pi/4+\varphi/2)+C \tag{10.63}$$

We use the convention that $\ln = \log_e$ indicating that this is the natural logarithm to base e. In the normal aspect cylindrical projections the origin of the plane coordinates is located somewhere on the equator. Therefore where $\varphi = 0$, $\mathrm{y} = 0$ and the integration constant, $C = 0$. Consequently the projection coordinates defining the Mercator projection of a sphere of unit radius are:

$$\mathrm{x} = \lambda \tag{10.64}$$

$$\mathrm{y} = \ln\tan(\pi/4+\varphi/2)$$

TABLE 10.04 *Mercator's projection. Values for the ordinate, particular scales and distortion characteristics for the 15° normal aspect graticule*

Latitude φ	Ordinate y	Particular scales h	Particular scales k	Area scale p	Maximum angular deformation $\omega°$
0°	0·0000	1·0000	1·0000	1·0000	0°
15°	0·2649	1·0353	1·0353	1·0719	0°
30°	0·5493	1·1547	1·1547	1·3333	0°
45°	0·8814	1·4142	1·4142	2·0000	0°
60°	1·3170	2·0000	2·0000	4·0000	0°
75°	2·0276	3·864	3·864	14·931	0°
90°	∞	∞	∞	–	–

In this case where the equator is a single line of zero distortion, the particular scales are

$$h = k = \sec\varphi \tag{10.65}$$

$$p = \sec^2\varphi \tag{10.66}$$

$$\omega = 0° \tag{10.67}$$

Numerical values for these are given in Table 10.04. The projection is illustrated in Fig. 10.10.

Because of the practical importance of this projection we must also consider the effect of the modification caused by the introduction of a standard parallel. This is frequently used for navigation charts which bear such statements as 'Scale 1/2 000 000 at 56°N'. If we denote this standard parallel by φ_0, then the particular scale in this latitude is

$$k_0 = \mathrm{dx}/[\mathrm{R}\,.\cos\varphi_0\,.\,\mathrm{d}\lambda] = 1 \tag{10.68}$$

or

$$= \mathrm{x}/[\mathrm{R}\,.\cos\varphi_0\,.\,\lambda] = 1 \tag{10.69}$$

and therefore

$$\mathrm{x} = \mathrm{R}\,.\cos\varphi_0\,.\,\lambda \tag{10.70}$$

Elsewhere on the map we have

$$k = \mathrm{R}\,.\cos\varphi_0/\mathrm{R}\,.\cos\varphi \tag{10.71}$$

so that the condition for a conformal projection must now be

$$\mathrm{dy}/\mathrm{R}\,.\,\mathrm{d}\varphi = \mathrm{R}\,.\cos\varphi_0/\mathrm{R}\,.\cos\varphi \tag{10.72}$$

From this we obtain the projection coordinates for the modified form of

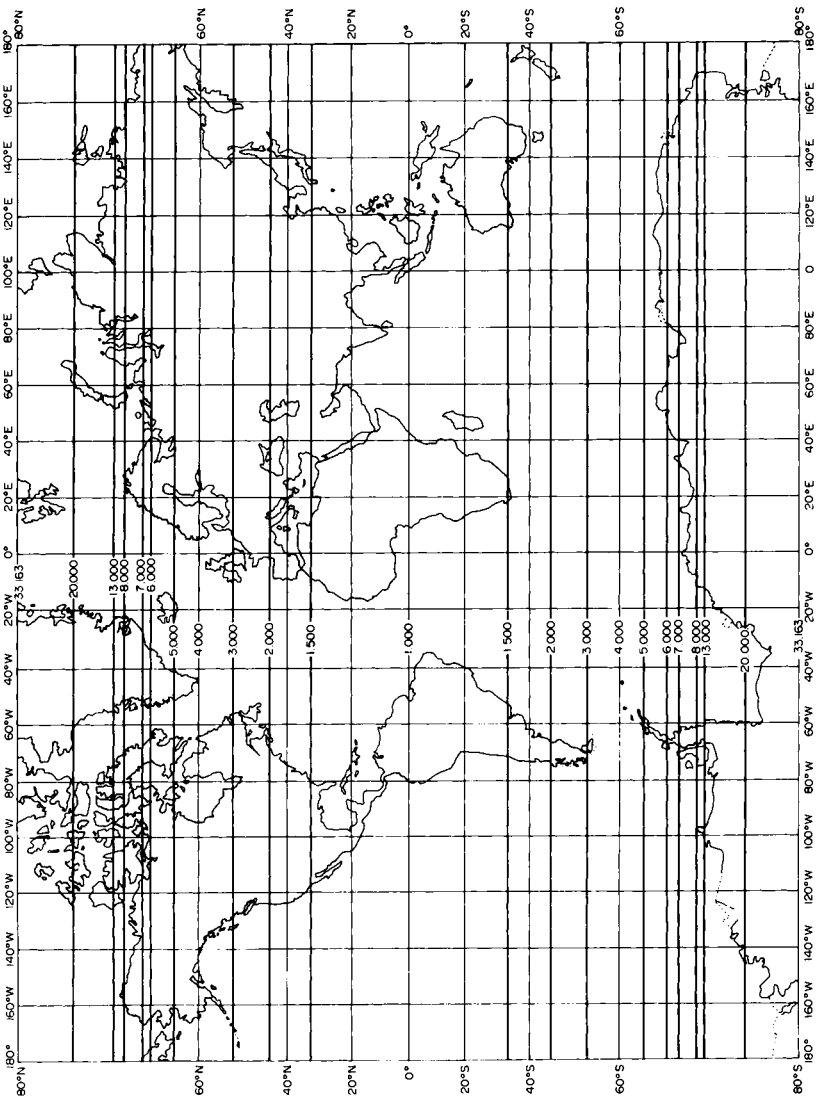


FIG. 10.10 The normal aspect of Mercator's projection. The isograms represent equal values of area scale (p). For illustration of other properties of Mercator's projection, see Figs 11.07, p. 237, 12.03, p. 250, 12.04, p. 252, 14.02, p. 297, 14.05, p. 300, 14.06, p. 303, 14.08, p. 306.

Mercator's projection

$$x = \cos\varphi_0 \,.\, \lambda \tag{10.73}$$
$$y = \cos\varphi_0 \,.\, \ln\tan[\pi/4 + \varphi/2]$$

Example IIIA: Derivation of Mercator's projection of the spheroid

The third stage in our study of this important projection is to show how the normal aspect Mercator projection can be derived for charts at scales where the spheroidal assumption is needed. From equation (4.12), p. 70, we already know that an infinitely short meridional arc on the spheroid may be expressed in the form

$$ds_m = \rho \,.\, d\varphi \tag{10.74}$$

and from equation (4.11) we have the corresponding expression for an infinitely short arc of the parallel

$$ds_p = \nu \,.\, \cos\varphi \,.\, d\lambda \tag{10.75}$$

We substitute these expressions in equations (10.53) and (10.54) defining the particular scales along the meridian and parallel. Thus

$$h = dy/\rho \,.\, d\varphi \tag{10.76}$$

$$k = dx/\nu \,.\, \cos\varphi \,.\, d\lambda \tag{10.77}$$

For the projection of the spheroid, equation (10.55) may be written in the form

$$x = a \,.\, \lambda \tag{10.78}$$

where a is the major semi-axis of the spheroid and, as in equation (4.10), p. 69, it is the radius of the equator. Therefore

$$k = a \,.\, d\lambda/\nu \,.\, \cos\varphi \,.\, d\lambda$$
$$= a/\nu \,.\, \cos\varphi \tag{10.79}$$

The condition for a conformal projection is now satisfied by the equation

$$dy/\rho \,.\, d\varphi = a/\nu \,.\, \cos\varphi \tag{10.80}$$

Substituting for ρ and ν their respective values according to equations (4.08) and (4.09), equation (10.80) becomes

$$dy = a\{[(1-e^2)d\varphi]/[(1-e^2\sin^2\varphi)\cos\varphi]\} \tag{10.81}$$

Integration of this equation leads to

$$y = a \,.\, \ln\tan(\pi/4 + \varphi/2)[(1-e \,.\, \sin\varphi)/(1+e \,.\, \sin\varphi)]^{e/2} + C \tag{10.82}$$

Here e represents the first eccentricity of the spheroid originally defined in equation (4.02), p. 65.

As in the derivation of Mercator's projection for the sphere, the integration constant $C = 0$. The part of the right-hand side of equation (10.82)

$$q = \ln \tan(\pi/4 + \varphi/2)[(1 - e \cdot \sin\varphi)/(1 + e \cdot \sin\varphi)]^{e/2} \qquad (10.83)$$

This is the *isometric latitude* referred to in Chapter 4. This is so-called because a system of (q, λ) coordinates upon the curved surface of the spheroid subdivides it into a network of small squares. The system of *isometric coordinates* thus defined may be employed to derive other conformal projections, and is therefore extremely useful in the further study of them.

A variety of different methods may be used to convert equation (10.82) into a form which is convenient for practical computation. The method commonly found in British and American works is to expand the term

$$[(1 - e \cdot \sin\varphi)/(1 + e \cdot \sin\varphi)]^{e/2}$$

as a series. This leads to the equation

$$y = a \cdot \ln \tan(\pi/4 + \varphi/2) - a[e^2 \cdot \sin\varphi + (e^4/3) \cdot \sin^3\varphi + (e^6/5) \cdot \sin^5\varphi + (e^8/7) \cdot \sin^7\varphi + \ldots] \qquad (10.84)$$

Values for the ordinate of Mercator's projection can usually be obtained from tables without having to calculate (10.63) for the sphere or (10.84) for the spheroid. For use with the spherical assumption there are numerical published tables of *Meridional Parts* (or *Mercatorial Parts*) because these are important in marine navigation. Since the abscissa of the Mercator projection varies only with longitude, the tables are usually compiled in arguments of minutes of longitude at the equator, giving the distance from the equator to any parallel φ, in these units of measurement. If a line to represent the equator is drawn and carefully subdivided at the required scale of a chart, the remainder of the construction can be done by setting compasses to the required separations given in tables of meridional parts. Cotter (1966), and textbooks on navigation, describe the technique in detail. Maling (1989), who deals with the special methods of correcting measurements made on Mercator's projection, treats also with various methods of calculating meridional parts using modern methods. Tables of meridional parts have been available since the sixteenth century. They are normally parts of more complete sets of tabulated mathematical functions which have been specially designed for ease of use in navigation. There are at least two sets of published tables of meridional parts for the spheroid, namely USHO (1932) and Hydrographic Department HP 470 (n.d.). Both of these are for the Clarke 1880 Figure of the Earth

($f = 1/293{\cdot}5$), which is appropriate for use in African waters, so that we must presume that, in the days before easier computing, the hydrographic charts produced in English-speaking countries were all based upon this reference spheroid. Today, of course, the meridional parts are so easily determined by pocket calculator that this historical oddity no longer matters.

CHAPTER 11

Choosing a suitable map projection – the principles

> Few people, even few cartographers, commonly know what projection is good for what purpose and the tradeoffs involved.
>
> P. Jankowski and T. Nyerges, *The American Cartographer*, 1989

Introduction

An infinite number of different map projections are theoretically possible. It is likely that only about 400 have been described and only about half of these have ever been constructed. Less than 50 of them have been commonly used, and excluding those used solely in atlases, fewer than 30 have been used for all purposes.

In most branches of cartography, notably in the preparation of large-scale maps, topographical maps and navigation charts, there is very little possibility of exercising any choice about the kind of projection to be used as the base for the map or chart. The most suitable projections for these purposes have evolved to meet the needs of the specialised user. Often, too, the specification of the projection has been adopted for use with related map series produced by the International Civil Aviation Organisation (ICAO), the North Atlantic Treaty Organisation (NATO) and others in an attempt to achieve some measure of standardisation.

In other kinds of cartographic work, especially in atlas production, there is a greater freedom of choice in selecting a projection which is suitable for a map of a particular country or continent and for a particular purpose. In this chapter we investigate some of the criteria and methods which may be used when it is possible to make this choice. Naturally this study is concerned primarily with the design and production of small-scale maps showing an entire country, a continent, a hemisphere or the whole world. In later chapters we will examine the practical reasons why only certain projections are preferred for use in navigation, surveying and topographical cartography.

Geographical and Land Information Systems

Thus far we have assumed that the projection is to be used for a conventional map, this being a map which has been drawn for reproduction on a sheet of paper. Today, however, a whole new field of cartography has developed through the implementation of geographical information systems (GIS). These comprise files of geographical or positional data stored in digital form, and the manipulation of such files in much the same way as we may use conventional maps. A land information system, or LIS, is to be regarded as being the equivalent tool for legal, administrative and economic decision-making and an aid for planning and development for much smaller areas. Although we must resist the temptation to embark upon an elaborate statement about the nature of such systems, there is still need to comment briefly about their nature and purpose. In the early 1990s there is still more discussion about what one day may be achieved by GIS, rather than concrete examples of what has actually been done. Indeed, Chorley (1988) has characterised a GIS as being *'a tool in search of a problem'*.

The salient features of a GIS may be represented diagrammatically in Fig. 11.01. The essential concept is that the system comprises a collection of digital files comprising positional data, all of which may be accessed by the system to unite data from disparate sources. For example at the level of the LIS for a municipality, the files may include information relating to the underground services of a town; water mains, sewers, gas pipes, electricity and telephone cables, which need to be matched with the surface detail of streets and buildings shown on the conventional large-scale maps, and possibly also with cadastral information relating to land ownership or tenure. At the national, or continental, level the GIS contains files of geology and soils, land use and vegetation maps, climatic data and the demographic, argicultural and other economic census returns for entire countries.

In the everyday work of surveyors, architects and planners, a well-known technique, in use for at least a century, has been to prepare transparent overlays to depict the different services so that these might be superimposed one upon another to show where one service is situated with respect to others, and might be used as a technique to control the actions of one group, for example the water engineers, from digging a hole which promptly damages the gas or electricity supply to an area. This process of preparing separate transparent overlays is replaced in a LIS by comparison of two or more files; for example, both gas and electricity services with the surface information about streets and buildings without first having to draw the overlays. The analogy is so close that the different subject files in a GIS or LIS are often referred to as *layers*.

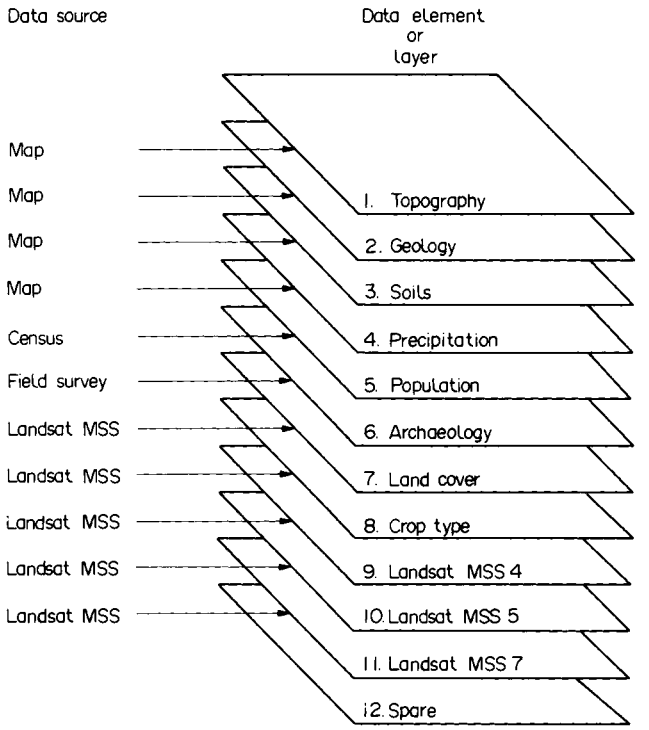


FIG. 11.01 Diagrammatic representation of a 12-level geographical information system. (Source: Curran, 1984.)

The subject of map projections enters the field of GIS in three ways. First, and in common with conventional cartography, it is necessary to decide how best to present the results of analyses, whether the output is an ephemeral display on the screen of a monitor or in the form of a printed map (called *hard copy*). Second, it may be necessary to reduce the contents of the different layers within a system to a common coordinate system before it is possible to match the data in a satisfactory manner. Third, it may be desirable to apply checks to any quantitative measurements made from the data contained within the system. The kinds of cartometric measurements which may be made internally are of distance, angle and area, which may then be combined with other data to create indices of density, gradient etc. Depending upon the nature of the projection used to hold the data, some form of correction ought to be made for the projection distortions which are inherent to the system. Some of the methods which may be applied have been described in Maling (1989) but, at present, the application of such corrections within a GIS is still very much in its infancy.

The role of the zero dimension in a GIS

We use the term *zero dimension* to describe the effective limit of what may be detected on a paper map with the naked eye, and which therefore represents a practical limit to uncertainty and errors in mapping, whether these arise from the original survey, or from the subsequent cartographic process, the influence of the map projection and subsequent cartometric work. We have already explained in Chapter 5 that at the larger scales, map sheets cover a relatively small area, and although the projection distortions are present they are too small to be measured. In other words they are smaller than the zero dimension.

Common experience of making and using maps sets the zero dimension at about 0·2 mm, which is the size of the finest point which is visible to the naked eye. Most cartographic draughting is about this order of precision, as has been described by the author in Maling (1989), although some writers, for example Tobler (1988), use a 'blunt pencil' criterion that the smallest physical mark which the cartographer can make is about one half-millimetre in size.

When maps were drawn only for reproduction on paper, some degree of generalisation was inevitable. Because many ground features are too small or too narrow to represent at their true scale size on a map, they have to be exaggerated so that these are legible and interpretable. Thus, as illustrated in Fig. 11.02, the *threshold of perception* must be matched by a *threshold of separation*, this being the smallest separation between symbols which still indicates that two separate objects on the ground are portrayed by two symbols on the map. It follows that if the threshold of separation is larger than the separation between features at map scale a small amount of exaggeration is introduced to the map. Because the feature now occupies more space on the map, other neighbouring information of lesser importance must either be deliberately shifted to a slightly different position or must be left off the map entirely. Huge significance lies in what the compiler of the map has regarded as being 'of lesser importance'. A map which has not been adequately generalised is usually an unreadable map.

Different standards of readability have come with digital processing of geographical information systems because of the ability of the computer to extract data from a file without reference to what the human eye can resolve. If the data files have been produced from large-scale maps (e.g. 1/2500 or 1/5000) the zero dimension corresponds to a much smaller limiting ground distance than if the source maps were of scale 1/25 000 or 1/50 000. Moreover, the size of the zero dimension changes with the kind of source material. For example, in the use of survey-quality aerial photography, the resolution of the camera lens, film and the optics of the plotter each amount to only a few micrometres. If we take the combined

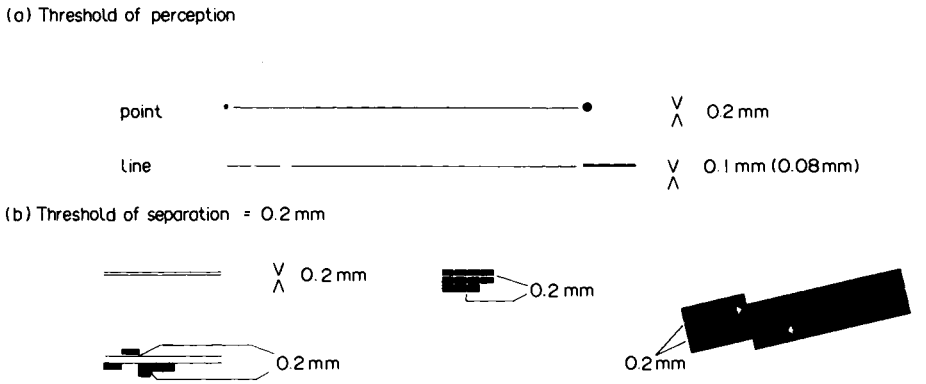


FIG. 11.02 (a) The threshold of perception and (b) the threshold of separation applied to map symbols. Both of these diagrams indicate that a threshold of about 0·2 mm applies, and that this is a reasonable value to take as the zero dimension. (Source: Rouleau, in ICA, 1984.)

effect to be about 15 μm for a diapositive viewed in a stereoplotter, this is a ten-fold improvement upon the 150 μm zero dimension of a map. In work with remotely sensed data acquired from satellites, such as the Landsat Thematic Mapper or SPOT HRV imagery it is possible to work to a zero dimension of only one pixel width. As a result the size of the zero dimension is much reduced.

Obviously the zero dimensions of each source are carried through to the corresponding GIS files. There is nothing magical in the digitising processes which can convert a discrepancy of 0·2 mm into a zero displacement. So, too, the small deformations due to the projections upon which the sources were based and which were hitherto small enough to be ignored because they could not be detected on a map. The existence of these residual discrepancies causes difficulties in computer matching of layers derived from different sources. This is a problem which the older generation of architects or planners did not experience, because all the overlays were equally crude tracings and visual interpretation of one distribution superimposed upon another could compensate for the *slivers and slices* created by small discrepancies along the boundaries. Computer processing locates and exhibits such features with unerring skill.

For example, in describing the limitations of some of the sources for databases with CORINE, the environmental GIS for the European Community, Briggs and Mounsey (1989) have described the problem of reconciling the data digitised from different sources. They have written:

> Possibly the most acute problems are likely to emerge when overlaying data sets (e.g. soils upon topography, or climate on soils). From experience to date, one of the most common failings of users is to misunderstand the limitations which map source scale imposes on these operations. Often, indeed, users believe that the database is inde-

> pendent of scale, due to the capability to plot or analyse data at any scale within the scope of the hardware. In practice, however, it is clear that data obtained from small-scale sources cannot realistically be analysed in conjunction with data derived from large-scale sources due to the inherent differences in accuracy and spatial precision. The only valid course is therefore to generalise the larger-scale data to be compatible with the smaller-scale data set. Whilst this will lead to some loss of information, in reality, of course, it is merely an admission of the relative in-built inaccuracies in the data sets.

The particular example which has created trouble in CORINE has been reconciliation of the detail on the 1/1 000 000 soils map of the EC and the topographic base originally derived from the 1/1 000 000 Operational Navigation Charts (ONC) and other positional data. In this respect Briggs and Mounsey state that:

> Although the soil map has been published as a single set of map sheets, the base maps on which they are drawn have been derived from topographic maps with different origins of projections, and this was 'fudged' in compiling the European map to ensure that a continent-wide continuous map could be assembled. This has resulted in distortion which, while negligible in the case of purely cartographic representation, produces significant discrepancies when attempts are made to overlay the results on other data sets... the experiences demonstrate that, even with apparently consistent datasets, considerable hidden inconsistencies may exist which need correction in constructing an operational, integrated information system. Further it demonstrates the dangers of the application of increasingly sophisticated tools within GIS software, without some feel for the original data, and some understanding of the nature of the results.

Since one of the sources of these discrepancies results from trying to combine data which have been mapped on different projections it follows that remapping of GIS sources to a single suitable projection is essential. Rather than repeat this operation every time a particular file is required, there is a need to choose a suitable projection to use in the GIS operating system in just the same way as this was formerly needed in compiling a series of thematic maps. Although the choice is likely to be one of the projections used for many of the sources, for example the Universal Transverse Mercator (UTM) projection (p. 357 *et seq.*) which is likely to be the projection used for the topographic base, the need is present to make this choice, or if other opinions prevail, to choose a suitable projection to be used within the GIS. This has to be done in the same way we would have to proceed to choose a suitable projection for a new atlas map. The nature of this requirement is considered further in Chapter 19, p. 408.

Some factors influencing the choice of a suitable projection

It is a fundamental principle of distortion theory that the particular scales, and therefore exaggeration, of areas and angles increase from the origin of the projection towards its edges. Since all projections have distortions of one kind or another and since, on a small-scale map showing a large

portion of the world, these distortions can be measured, it is usually desirable to choose a projection in which distortion is tolerably small. Thus the primary aim of a logical choice is *to select a projection in which the extreme distortions are smaller than would occur in any other projection used to map the same area.* We shall see in Chapter 13, pp. 281–289, that sometimes the converse argument is used and a projection may be chosen because it deliberately exaggerates some feature or some part of the map. This may be done to assist the tourist as, for example, in the variable scale town maps pioneered by Falk Verlag. In scientific applications the concept is used to collapse or extend space, and thus illuminate distributions which would otherwise be too dense or too sparse to interpret if plotted on a conventional map. Common use of such techniques dates from 1957 when Hagerstrand used the *logarithmic azimuthal projection* (p. 282) to illustrate migration from a rural community in Sweden.

The amount of distortion which is likely to be encountered in a conventional map depends upon the location, size and shape of the area to be mapped. Distortion is least in the representation of a small, compact country and greatest in maps of the whole world. The three variables – location, size and shape – usually determine the choice of origin, aspect and class of a suitable projection.

The purpose of the map and its intended use

The purpose of the map, especially a certain knowledge of the ways in which it is going to be used, generally determines which special property is important. For example, if we need a conformal map of a country, we may study the way in which the area scale increases near the boundaries of the country and select that conformal projection which shows the least exaggeration of area within the parts to be mapped. If we require an equal-area map of the country, we must carry out a similar evaluation of the angular deformation inherent to all equivalent projections. If neither special property is essential, examination of both area scale and angular deformation must be made. This kind of evaluation suggests that the concept of minimum-error representation, briefly mentioned in Chapter 6 as a special property, may be valuable in this context. Consequently we proceed from the hypothesis that 'the best projection for a country' is likely to be the minimum-error projection which also satisfies another special property which is also deemed necessary in map use. We shall find that the requirement for a special property is most exacting in the design of navigation charts, and in the relationship between surveying and quantitative map use, but much less so for other uses. It is therefore suggested that the expected quantitative uses of a map or chart, to measure distances, areas and angles, are more likely to expose the inadequacies of

a projection than are any subjective visual methods of appraisal. Therefore we consider desirable criteria in these terms.

The present author has shown, in Maling (1968a), Frolov and Maling (1969) and Maling (1989), that it is reasonable to expect measurements of distance or area made on a map to have relative precision of the order of $\pm 1\%$ to $\pm 2\%$ provided that reasonable precautions have been made in using the appropriate instruments. For many purposes *other than* navigation, artillery, surveying and some engineering applications, angles measured on maps are usually not needed with an accuracy greater than 1°. These criteria have been tentatively recommended as indicating the tolerable amounts of deformation acceptable in the projection to be used for a map in a national atlas, but the extent to which they can be satisfied depends upon the location, shape and size of an individual country.

Maps of small areas or small countries

If we are required to select a suitable projection to depict a small, compact country, and we are free to choose any point on the earth's surface as the origin for the projection, then the possibilites are practically limitless, or, in other words, it matters little which projection is used. In all classes of map projection the distortion in the vicinity of the point or line of zero distortion is less than the zero dimension, so that it cannot be detected by measurement on the map. In studies using conventional maps the influence of the projection is generally ignored. This is equivalent to selecting the projection which has been used as the base of the topographic map. Thus a geographer or planner wishing to produce a distribution map of part of England at a scale of 1/500 000 or larger would not be preoccupied with the merits of which projection to use, but would plot the new information on existing Ordnance Survey sources. Ordnance Survey maps of scale 1/625 000 and larger are all based upon a version of the *Transverse Mercator projection* (pp. 354–356), which is conformal, but the amount of exaggeration in area which is introduced by using this rather than a truly equal-area map is trivial. The area scale on this version of the projection nowhere exceeds the range 0·99908–1·00092 in mainland Britain, i.e. it varies from the constant area scale of an equal-area projection by less than 0·1%. Consequently judgements about density of distribution or measurement of area occupied by different categories of land use, for example, are unaffected by the fact that the map projection used is theoretically incorrect.

Maps of large countries, continents and the whole world

Just as the choice of a suitable projection is unimportant in the design of a distribution map of mainland Britain so, too, most of the individual

countries of western Europe can be adequately represented by using the national projections adopted for topographical map series. A map of the whole of western Europe can be prepared without exceeding the $\pm 2\%$ and 1° tolerances which have been suggested. By contrast it would be difficult to find a projection to map the whole of Canada or the USSR in which linear or area distortion is less than 3% or angular deformation is less than $3–5^\circ$. For maps to represent entire continents or oceans much larger amounts of deformation must be tolerated. For example, an equal-area map of Asia involves the presence of maximum angular distortion of about 15° somewhere near the edges. Equal-area maps of Africa and North America have maximum values for ω in the range $6–8^\circ$. Figure 12.01, p. 247, illustrates this for a map of North America, and Fig. 12.02 for a map of China. Equal-area maps of the hemispheres show an increase in ω to about 30°. Map projections of the whole world generally have singular points where $\omega = 180^\circ$ and p is indeterminate, but even if these extreme values are discounted as being inevitable and therefore unrealistic measures of the remainder of the map, we must expect to find that angular deformation greater than 45° or area scales in excess of 2·5 (+250%) must be tolerated in some parts of the map. Then the real skill in selecting a suitable projection is to arrange for the important parts of the world map to lie where the distortions are least. This leads us to a consideration of the intended purpose of the map and the extent or nature of the distribution which is to be mapped.

Modified projections

The use of the word *modification* when applied to a map projection suggests a wide variety of possibilities. For example, one might argue, with a certain justification, that the change in the appearance of the graticule with change of aspect should also be called modification. The following four methods might reasonably be understood to represent modification of a projection, though the author has argued, in Maling (1968b), that it is preferable to retain the word modified to describe only the first of these.

- modification through redistribution of the particular scales and the creation of more than one line of zero distortion;
- modification through the introduction of special boundary conditions on the edge of the map;
- transformation by repetition of part of a map projection giving rise to a *recentred* or *interrupted* projection.
- transformation through the combination of different map projections to give an appearance of continuity of the map.

Several of these techniques may be used in the same map, particularly

for world maps where the problems of distortion are obviously most pronounced.

Obstacles to choice

In contrast to those factors which must be considered to influence the choice of a projection for a new map there are some practical obstacles which limit freedom of choice. Most of these are owing to the cost in time and labour of compiling, plotting and redrawing maps on projections which differ from the sources used. We have already seen in Chapter 8 that the creation of a new map projection may involve some exceedingly careful plotting and drawing if this has to be done manually. Yet the completion of this stage of the work is only a preliminary to the extremely slow job of transferring map detail within the new graticule.

Excluding the use of digital mapping methods, there is no quick or simple optical method of transforming map detail in one step which is comparable to the use of the process camera for changing scale. Optical *rectification*, similar to the procedures used in photogrammetric mapping, has been used in some establishments where photogrammetric rectifiers, such as the old Zeiss SEG 1 instrument, may still be used for this purpose. Special *optical pantographs*, like the Grant projector and Röst Plan Variograph, can be obtained with tilting easels which similarly permit partial rectification of the source map to fit the new graticule. However, the range of these applications is quite limited, for optical rectification cannot transform a rectilinear graticule into one comprising families of curves, or vice-versa. Thus we cannot transform a normal aspect cylindrical projection into a normal aspect conical projection, and the only change we can make to the original rectilinear graticule is to transform the rectangular quadrangle into a trapeziform rectilinear figure (polyhedric projection) or to alter the ellipticity of the elliptical meridians of a Mollweide or Hammer–Aitoff graticule. More elaborate optical–mechanical apparatus has occasionally been designed for specific purposes. For example, Honick (1967) described equipment used for transforming the graticules of aeronautical charts, and in so doing demonstrated that an analogue solution to the problem can be rather complicated. Similarly in the days before digital mapping it required a series of photographs of a map mounted on a curved surface to make the source maps needed to produce the variable scale town maps based upon the *hyperboloid projection* (p. 283).

The manual work of plotting usually has to be done point by point after drawing a close network of corresponding geometrical figures on both source map and plotting sheet and transferring the map detail manually with reference to these lines. In this respect it is easier to transform from one conformal projection to another rather than to trans-

form to equal-area or other projections. This is because a very small part of a conformal map corresponds in shape to that part of the map being compiled. Thus, as O. M. Miller (1941) has written,

> Of the two evils, the cartographer dislikes the conformal type of projection less, because he knows that, provided he makes the mesh of his grid small enough, detail, if properly reduced by pantograph or photography, will fit nicely into place and can be traced directly on the map being compiled.

Because plotting and redrawing of the detail may represent many weeks or months of work it is not surprising that changing the projection of a map was always commercially unpopular. The first reaction of many cartographic editors to such a proposal was to consider whether there was any existing material which was suitable for use for a particular map in a new atlas. Robinson (1952) castigated the commercial map producer who was 'only too happy to peddle the older wares', and it is easy to condemn the reissue of atlas maps as exemplifying lack of initiative or new ideas. However, the bleak commercial fact remains that the jobs of compilation, plotting and drawing are the most expensive stages in conventional map production, and these are essential for the production of a new map on a different projection. It was nearly always cheaper to revise existing fair drawings.

The digital solution is the most successful method of overcoming these difficulties because, as described elsewhere in this book, it is possible to transform information which has been digitised or scanned from the form in which it is stored into the master grid coordinates and plot the map at the required scale. Moreover, the whole of the process of compilation and fair-drawing can be executed with a minimum of human interference. However, the method depends entirely upon the availability of a suitable GIS layer comprising the map detail in machine-readable form. The acquisition of such databases by digitising is also slow and expensive.

To the pioneers of digital mapping the grand design was that of the *cartographic databank*, based upon digitising the basic scale mapping, that is the largest scale maps of each country so that the information contained in the system would be least affected by the generalisation which characterises smaller-scale maps.

Inevitably there were formidable practical difficulties to be overcome, in the acquisition of such databanks; for example the time needed to digitise the source maps and store the results in adequate and accessible form. Attempts to create such massive collections of data soon led to the realisation that the paper map was a far more compact way of storing positional information than was possible in any existing computer hardware and, indeed, this was true until the middle 1980s when optical disc technology entered this field. Only now is it possible to store the topographical map cover of a country economically in the form of CD-ROM discs (compact disc read-only memory). Nevertheless, at present,

and for decades to come, the available databanks are still restricted in extent, content and utility to small-scale cover of a country.

Thus the *scale-free* databases, which correspond to the original databank concept and are so-called because they have been created from the largest available *basic scale* mapping, are still largely confined to the English West Midlands and a few other scattered blocks of urban mapping. According to Proctor (1986) the completion date for digital coverage of England, Scotland and Wales by the Ordnance Survey is 2015, which is too far distant for most user needs in the 1990s.

Databases for the whole world are, of necessity, still extremely generalised; for a brief note on these see Tomlinson (1988). There are, for example, two world databases now in the public domain, which were originally prepared in the early 1970s by the Central Intelligence Agency. World Database I was prepared from source maps of scale 1/12 000 000, and therefore extremely generalised. For example it was used for the coastlines of the majority of the maps in Snyder and Voxland (1989). World Database II was prepared from sources at scale 1/3 000 000 or thereabouts. It follows that although digital methods provide a wonderful opportunity for experiment in using other projections, the lack of availability of data will remain an obstacle to progress for some decades to come.

The choice of origin, aspect and class of a projection

The preliminary stage in making the choice of a projection is to consider the location of the origin. In order to avoid excessive distortion within the area to be mapped, we locate the point or line of zero distortion near the centre of it and orientate the lines of zero distortion to the longer axis through the country. This choice of origin and orientation of the lines automatically affects the aspect of the projection. The shape of the area to be mapped influences the choice whether it should be a point or line of zero distortion and this, in turn, determines the class of projection. Thus all three variables are intimately related and must be considered together.

The traditional approach to the choice of class is described in most of the elementary textbooks by the following three rules.

- if the country to be mapped lies in the tropics, a cylindrical projection should be used;
- if the country to be mapped lies in temperate latitudes, a conical projection should be used;
- if the map is required to show one of the polar regions, an azimuthal projection should be used.

These rules follow logically from the fundamental properties that the

principal scale is preserved along the equator in a normal aspect cylindrical projection, along a parallel of latitude in a normal aspect conical projection and at the geographical pole of a normal aspect azimuthal projection. The principles have been applied to the design of most sheet and atlas maps published since the sixteenth century; indeed they may be regarded as being *one of the classical foundations of cartographic design.* However, these should not be regarded as being inflexible rules. After all, no mention has been made of any of the other named classes of map projections, and these also deserve consideration in making the choice. Moreover, strict adherence to the three principles ignores the considerable advantages to be gained from using a map projection in any of its other aspects. In other words, the three rules are too restricting to be rigidly applied in modern cartography. For example, Fig. 10.02 shows that the normal aspect Azimuthal equal-area projection is a useful base for distribution maps of the Arctic Ocean or Antarctica, conforming to the third rule given above. But the transverse aspect (Fig. 10.03) of the same projection would be equally valuable as the base for a map of the Indian Ocean and the simple oblique aspect (Fig. 10.04) of it for mapping distributions of the North Atlantic Ocean. The use of an oblique aspect azimuthal projection is no longer to be regarded as a novelty. Transverse and oblique cylindrical projections are well known in large-scale and topographical cartography, but are much less often used for atlas maps. Much rarer are the transverse and oblique aspect conical projections. The *Bipolar oblique conformal conical projection*, Fig. 11.03, designed by O. M. Miller and W. Briesemeister for the American Geographical Society in 1941, is one of the few examples of oblique aspect conical projections which have become well known. In Chapter 12 we use this classic study to find a projection suitable for a general reference map of Hispanic America as an example of the combined graphical and analytical approach to choice.

Since we are able to select any point on the earth's surface as the origin of a projection, we may locate this at or near the centre of the country or continent to be shown on the map. The point of origin might be determined by computation, for example, as the centre of gravity of the land mass, using the standard methods of calculating this for a plane figure shown by the outline of the country or continent on any convenient

FIG. 11.03 Map of the Americas on the Bipolar Oblique Conformal Conical projection. The isograms represent equal values of linear deformation of −3·5%, 0%, +3·5% and +10%, corresponding to the particular scales of 0·965, 1·000, 1·035 and 1·100, respectively. Note that the graticule on the map is composed of parallels at intervals of 4° in latitude and the meridians are shown at 6° intervals of longitude (at intervals of 12°, north of 60° North). This graticule corresponds to the system of sheet lines adopted for the International Map of the World (IMW) at scale 1/1 000 000. (Source: Miller, 1941.)

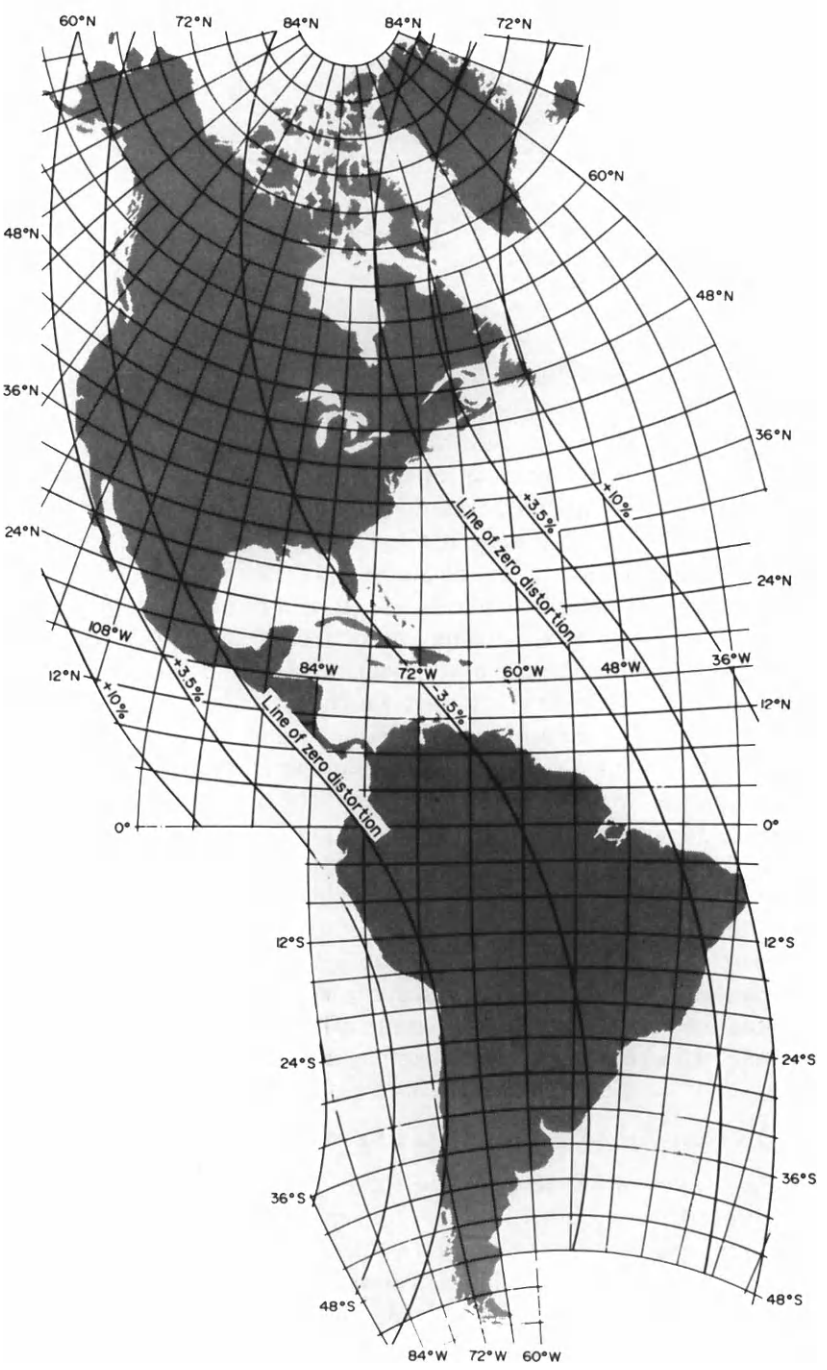

60°N
72°N
84°N
84°N
72°N
60°N
48°N
48°N
36°N
36°N
24°N
24°N
108°W
12°N
84°W
72°W
60°W
48°W
36°W
12°N
0°
0°
12°S
12°S
24°S
24°S
36°S
36°S
48°S
48°S
84°W
72°W
60°W
+3.5%
+10%
Line of zero distortion
+3.5%
+10%
Line of zero distortion
-3.5%

TABLE 11.01 *Suggested positions for the points of origin for maps of the continents*

	φ_0	λ_0
Europe	+50°	+20°
Asia	+40°	+95°
Eurasia	+40°	+85°
Africa	0	+20°
North America	+45°	−95°
South America	−20°	−60°
Australia	−25°	+135°

map. The method will almost certainly locate the origin at a point which does not correspond to any graticule intersection required on the finished map. The choice has to be made whether to calculate the projection with reference to this origin or to select the graticule intersection nearest to this point as the origin. Using modern computing methods there is no really great problem either way, for it is as easy with a pocket calculator to access the sines and cosines of an angle of 57°18′25″ as it is for 55°, whereas in the days when we had to use (z, α) tables, it was necessary to work from an origin at the nearest tabulated value for φ_0. This might differ from the required centre by as much as $2\frac{1}{2}°$ in latitude and longitude. Table 11.01 lists some of the points which might have to serve as the origins for maps of the continents working with this 5° module.

Usually the line of zero distortion is made to coincide with the longer axis through the country, or a pair of lines if the country is asymmetrical, like Chile, Japan or Indonesia. For example a map of Chile may be based upon a transverse cylindrical projection because the longer axis is practically meridional. On the other hand, maps of Japan and Indonesia require the use of an oblique aspect cylindrical or conical projection. Hammer (1889) illustrated the use of oblique aspect conical projections for Japan and South America a century ago. In Chapter 12 we shall investigate the suitability of an oblique aspect conical projection for a map of Latin America, although this may not be apparent at first sight.

Young's Rule for selecting class of projection

The choice to be made between the three classes of cylindrical, conical and azimuthal projections may be conveniently described in terms of *Young's Rule*, originally stated by Young (1920) and independently discovered and further extended by Ginzburg and Salmanova (1957).

The principle arises from the basic idea that a country which is approximately circular in outline is better represented by means of one of the azimuthal projections, in which distortion increases radially in all direc-

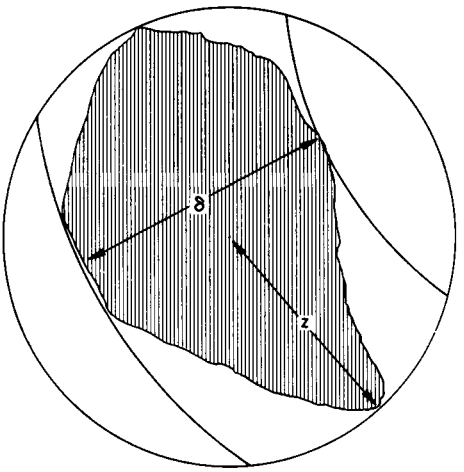


FIG. 11.04 The application of Young's Rule to the choice of a suitable class of map projection for a country with maximum extent z and minimum width δ.

tions, whereas an asymmetrical country is better mapped on a conical or cylindrical projection with lines of zero distortion.

The rule may therefore be described in terms of an imaginary country illustrated in Fig. 11.04. The area to be mapped has maximum angular distance z from the centre of the country to its most distant boundary. It can also be regarded as being bounded by two parallel arcs of small circles which lie $\delta°$ apart. These small circles may be parallels of latitude if the greatest extent of the country is east–west but, as implied by the orientation of these lines in Fig. 11.05, this is not a necessary condition of definition. Since we are concerned with the comparison of the particular scales and the distortion characteristics to be derived from them, we choose the pair of small circles which are the closest which can be fitted to this outline irrespective of their orientation to the conventional graticule. Note that we are going to compare the *maximum* radial distance, z, with the *minimum* separation of parallel circles, δ.

Young originally noted that if $z/\delta < 1{\cdot}41$, an azimuthal projection is to be preferred. Conversely if z/δ is greater than this critical value a conical projection should be used. Ginzburg and Salmanova have obtained three different critical values for z/δ depending upon the special property. From their study of the variations in particular scale in the ranges $0 < z < 25°$ and $0° < \delta < 35°$, together with the extension of the method to include cylindrical projections, these are

Conformal projections	$z/\delta = 1{\cdot}41$
Equidistant projections	$z/\delta = 1{\cdot}73$
Equal-area projections	$z/\delta = 2{\cdot}00$

The following examples are instructive. In Chile the total extent in latitude is approximately 32° but the greatest extent in longitude is only 7°. Hence putting $z = 16°$, $\delta = 7°$ we find $z/\delta = 2{\cdot}3$. This indicates that a conical or cylindrical projection is more suitable than an azimuthal projection and, as we have already seen, the best choice is a transverse cylindrical projection. For Australia the corresponding values are $z = 19°$, $\delta = 30°$ and $z/\delta = 0.63$. This indicates a preference for an azimuthal projection, which was the conclusion also reached by Sear (1967) in his valuable account of the arguments used to select the projection for a general reference map of Australia.

Choice of special property

We have already noted that the choice of special property is largely determined by the intended purpose of the map. In atlas cartography the special property of equivalence is especially important for mapping statistical data. However, it would be wrong to imagine that all maps in world, regional or national atlases are multipurpose maps for reference purposes. Since these are not necessarily intended to demonstrate density of distribution through clustering of dots, or for area measurement purposes, there is no particular reason why they should be rigorously equivalent. Since conformality and equivalence are mutually exclusive special properties, it follows that the exaggeration of area on a conformal map tends to be large, and that the angular deformation on an equal-area map also tends to be large. Between these two properties, which for practical purposes may be regarded as being the two limits of choice, there are a variety of other map projections in which neither property is satisfied, but they do not have the large distortions which are characteristic of conformal and equal-area maps.

We may demonstrate this by comparing area scale and maximum angular deformation for members of the azimuthal, conical and cylindrical classes of projection with the ranges $0° < z < 25°$ and $0° < \delta < 35°$, appropriate for maps of large countries. These are represented graphically in Figs 11.05, 11.06 and 11.07. From these graphs we see that, for azimuthal projections, the area scale of the Stereographic is approximately three times greater than the corresponding values for the Azimuthal equidistant projection, and the maximum angular deformation for the Azimuthal equal-area projection is appreciably greater than that for the Azimuthal equidistant projection. In conical and cylindrical projections the area scales of conformal maps are about twice as large as the corresponding values for the equidistant projections. The angular deformations of equal-area conical and cylindrical projections are approximately twice as large as for the equidistant versions. This leads us to the conclusion, already noted in Chapter 6, that the property

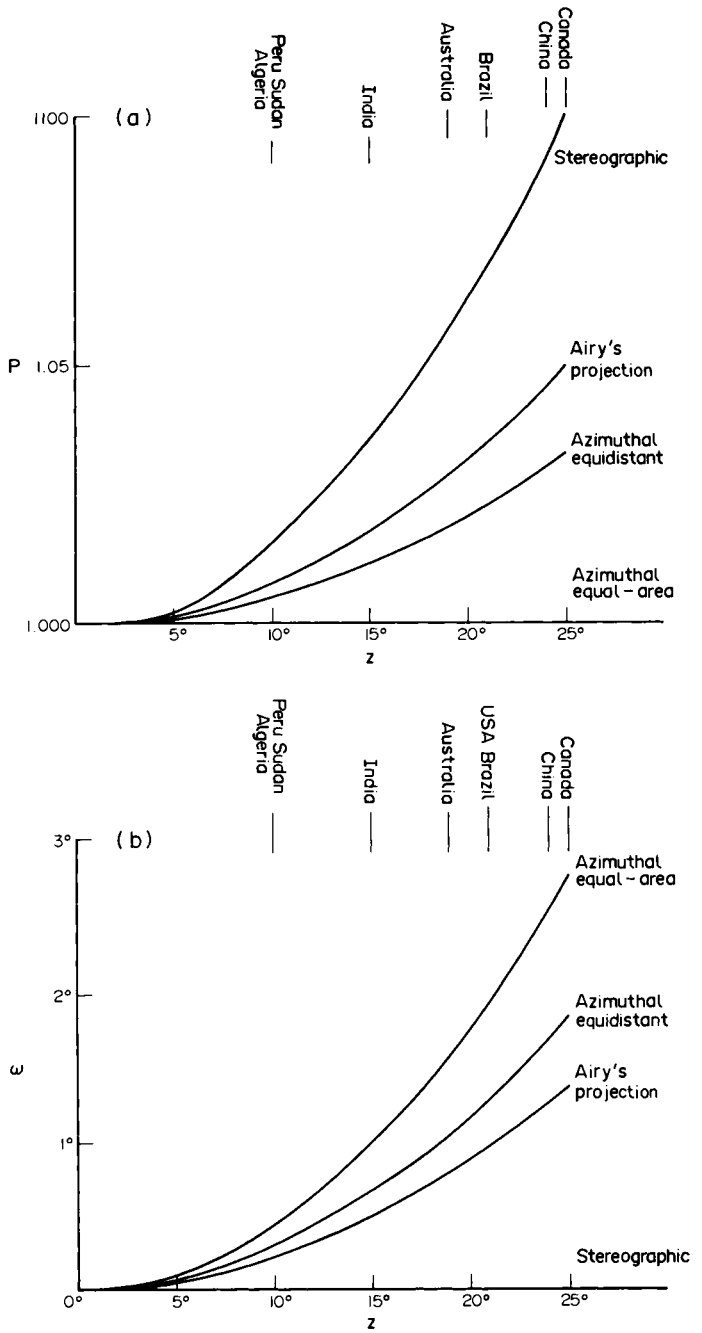
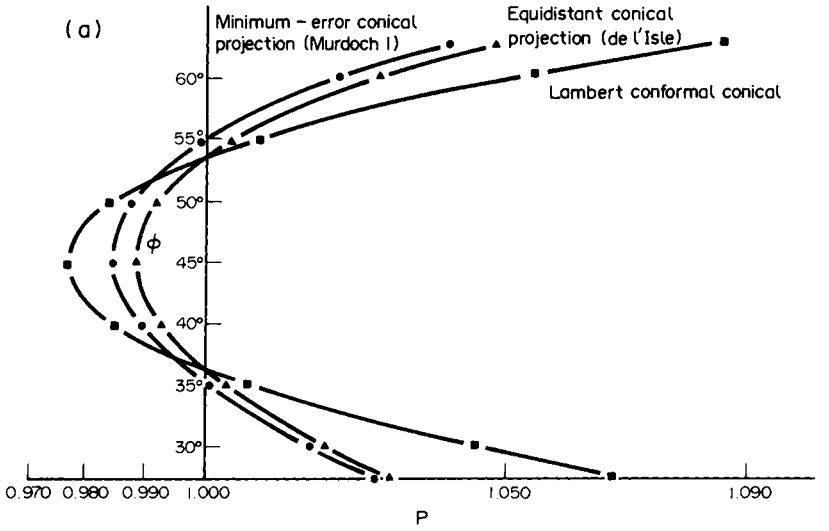


FIG. 11.05 The distortion characteristics of certain azimuthal projections within the range $0° < z < 25°$: (a) illustrates area scale (p) plotted against angular distance (z); (b) illustrates maximum angular deformation (ω) plotted against angular distance (z). The diagram also shows the approximate extent of certain countries according to the definition of z illustrated by Fig. 11.04.



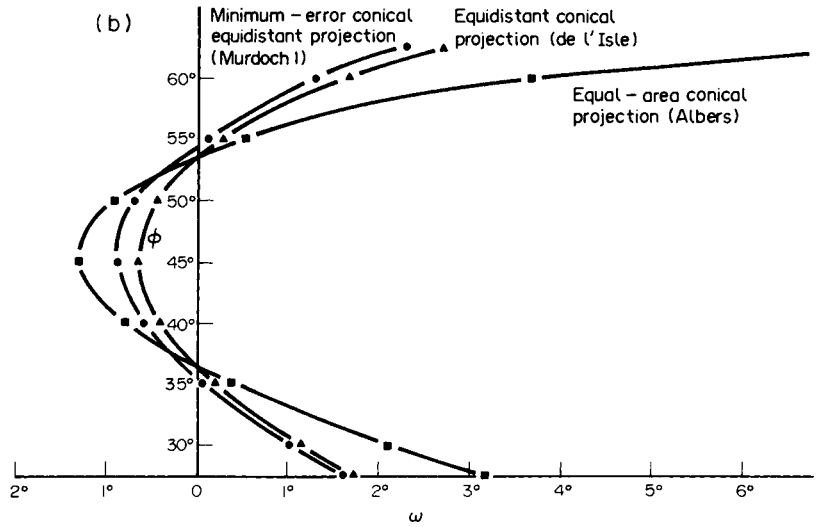


FIG. 11.06 The distortion characteristics of certain conical projections within the range $0° < \delta < 35°$ assuming the normal aspect and that the middle parallel corresponds to latitude 45°: (a) illustrates area scale (p) plotted against latitude (φ); (b) illustrates maximum angular deformation (ω) plotted against latitude (φ).

of equidistance often provides a useful compromise for use in maps which do not necessarily have to be rigorously equivalent or conformal. Hence we may regard the equidistant projection as occupying the central position within the continuum of all map projections listed by special property as

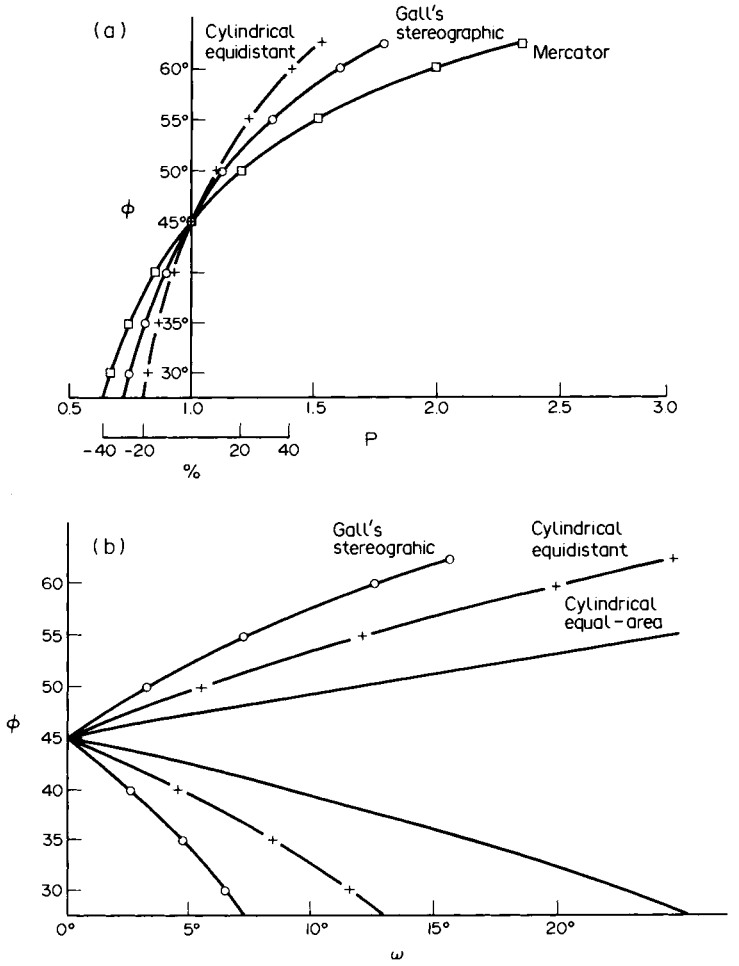


FIG. 11.07 The distortion characteristics of certain cylindrical projections within the range $0° < \delta < 35°$ assuming the normal aspect and that the line of zero distortion lies in latitude 45°: (a) illustrates area scale (p) plotted against latitude (φ); (b) illustrates maximum angular deformation (ω) plotted against latitude (φ).

illustrated in Table 11.02. There is also a close relationship between the distortion characteristics of equidistant and minimum error projections of the same class. This is shown in Table 11.03 by comparison of the distortion characteristics of the Azimuthal equidistant projection and Airy's projection within the range $0° < z < 25°$.

The mathematical theory of minimum-error representation was studied in some detail by Young (1920), and more recently by Snyder (1985), who both start from the same initial premise as Airy and Clarke, that the sums

TABLE 11.02

		Special property	Main uses
↑	↑	Conformal projections	Navigation charts, topographical, military and large-scale maps, synoptic meteorological charts
		(Projections with small angular deformation)	Small-scale strategic planning maps Climatic and oceanographic distribution maps.
−	+		
ω	p		
+	−	Equidistant projections	General reference maps
		(Projections with small exaggeration of area)	Atlas maps
↓		Equivalent projections	Statistical distribution maps

of the squares of the scale errors integrated throughout the area of the required map should have a minimum value. This was indicated by equation (6.33) on p. 109, which the reader will now appreciate is the form of expression needed to derive a minimum-error azimuthal projection. We further remind the reader that the concept of minimum-error is not an exclusive special property. Thus we may create minimum-error conformal or minimum-error equidistant projections which retain the special property, together with the additional advantage that the sums of square of the scale errors within the area to be mapped are less than in the parent projection. This is generally obtained through the modification of the parent projection by means of a suitable scale factor. We return to this problem in the next section. Following our preoccupation throughout this chapter with the need to reduce distortion towards the edges of a

TABLE 11.03 *Distortion characteristics (ω and p) for the azimuthal equidistant projection and Airy's minimum-error projection*

	Area scale (p)		Maximum angular deformation (ω)	
z	Equidistant	Airy's	Equidistant	Airy's
0	1·0000	1·0000	0°	0°
5	1·0013	1·0019	0°04′	0°03′
10	1·0051	1·0077	0°17′	0°13′
15	1·0115	1·0174	0°39′	0°29′
20	1·0206	1·0313	1°10′	0°52′
25	1·0325	1·0496	1°50′	1°21′

map, together with the suggestion that many general reference maps do not have to satisfy conformality or equivalence, it might be assumed that the correct choice of projection which best fits a given country is always the minimum-error projection of the selected class. Theoretically this conclusion is generally correct, but, in practice, the use of minimum-error projections is the exception rather than the rule. Consequently we are able to quote only four examples of the use of them in British and Commonwealth cartography during the twentieth century. These are:

- *The Ordnance Survey Ten-Mile map* (1/633 600) of the British Isles published between 1903 and 1936. This was based on *Airy's projection.*
- The use of a version of *Clarke's minimum-error perspective azimuthal projections* as the base for the synoptic meteorological charts published by the Meteorological Office in the Daily Weather Report. Use of this projection was discontinued in 1955 when it was replaced by the Azimuthal equidistant projection. In 1964 this, in turn, was replaced by the Stereographic projection.
- Hinks's (1942) choice of a *Minimum-error conical projection* (*Murdoch's third projection*) for the *British Council Map of Europe and the Near East* (1/11 000 000) published by the Royal Geographical Society in 1942.
- Sear's (1967) choice of the *Minimum-error azimuthal equidistant projection* for the map of Australia at 1/6 000 000, published by the Commonwealth Division of National Mapping in 1956.

There are probably two reasons why such little use has been made of minimum-error projections. First, the mathematical theory of minimum-error representation is difficult. Secondly, the primary source on this subject was, until recently, a booklet which never had a wide circulation, published nearly 70 years ago. As a result neither the theory nor the terminology are commonly known to cartographers and map users. Thus *Airy's projection* and *Murdoch's third projection* are seldom used, whereas the Azimuthal equidistant and Conical equidistant projections occur often in atlases. Table 11.03 has indicated that within the range of z which is needed to map most large countries, and even some of the continents, the differences between ω and p which exist between the little-known Airy's projection and the well-known Azimuthal equidistant projection are trivial. Although the mathematically correct answer to the question: 'What is the best map projection to use for a particular country?' is usually 'The minimum-error projection of the most suitable class', in practical cartography the equidistant projection of that class will provide a very similar map.

With the greater freedom and flexibility allowed by digital mapping methods, the mathematical and production constraints which were such

formidable obstacles to cartographic innovation a generation ago have been greatly diminished. If it is possible to design and redraw new maps and reproduce them as hard copy at little extra cost than to reproduce those already existing, there is greater encouragement to try new methods.

Modification through redistribution of the particular scales

In the brief accounts of the fundamental properties of the azimuthal, cylindrical and conical projections we have referred these to the tangent plane, cylinder or cone, but have not considered the alternative geometrical concepts illustrated by Figs 5.08, 5.09 and 5.10 on pp. 91–92. There it was shown that the effect of making the plane, cylinder or cone intersect the spherical surface is to replace the single line of zero distortion by two such lines, or to substitute a *standard circle* for the single point of zero distortion. We now investigate the significance of these changes.

On a conical or cylindrical map projection with a single line of zero distortion the particular scales increase outwards from this line towards the edges of the map. This is exemplified by the numerical values for the maximum and minimum particular scales for the equidistant conical projections given in Tables 10.02 and 10.03, pp. 207 and 210. If the single line of zero distortion of a conical projection is replaced by two *standard parallels* the effects upon the particular scales are as follows:

1. *Between the standard parallels and the edges of the map* the relationship between the maximum and minimum particular scales is similar to that for the unmodified projection. Thus in all normal aspect cylindrical equal-area projections the particular scale along the parallel is maximum and that along the meridian is minimum.
2. *The principal scale is preserved on both standard parallels.*
3. *Between the two standard parallels* the directions of maximum and minimum particular scales are reversed. Thus, in the de l'Isle projection, the particular scale along the meridian is maximum and that along the parallel is minimum. The following features should be noted:
   - *modification should have no effect upon any special property of a projection* – thus the de l'Isle projection is also equidistant;
   - *modification by the introduction of two standard parallels reduces the deformations towards the edges of the map* – we see in Table 10.03 that the maximum angular deformation in latitude 75° is 7°, whereas the corresponding value from Table 10.02 is more than 15°;
   - *modification has no effect whatsoever at the singular points* – for example in all normal aspect cylindrical projections the geographical poles are singular points where distortion theory is

invalid. Consequently the numerical values for $\varphi = 90°$ in both Tables 10.02 and 10.03 do no more than indicate this fact.

Precisely the same reasoning may be applied to cylindrical projections. Modification of this sort naturally has some effect upon the appearance of a projection. In the normal cylindrical projections the ratio between the length of the equator and that of a meridian is changed by the choice of standard parallel. The actual ratio depends upon the special property.

Modification of conformal projections is especially easy to apply because the particular scales are the same in all directions. This follows from the definition of conformality by equation (6.26) on p. 106. It is therefore possible to transform the coordinates of points and obtain the particular scales by using a single numerical constant or *scale factor*, as a common multiplier. The numerical value of the scale factor represents the particular scale to be preserved where the line of zero distortion is located on the unmodified projection. The value of it is related to the positions of the two lines of zero distortion so that a change in one results in alteration of the other. This kind of modification is commonly used with the varieties of conformal projection (*Transverse Mercator projection* and *Lambert Conformal Conical projection*) which are used in surveying and topographical cartography, as described later in Chapters 15 and 16, pp. 310–363.

The choice of standard parallels

Since we have established that there are advantages to be gained from redistributing the particular scales by means of standard parallels, it is desirable to consider how best to choose suitable standard parallels.

In equations (10.42)–(10.52) (pp. 208–209) we derived the equations for the Conical equidistant projection (de l'Isle) with two standard parallels and used the simple expedient in (10.42) and (10.43) of locating these midway between the central and limiting parallels of the zone to be mapped. This is, in fact, how the de l'Isle projection ought to be defined, to distinguish it from *Euler's projection* and the other equidistant conical projections which may also be described if we specify that certain ratios must be maintained between the bounding parallels and one near the middle of the map. A detailed account of the various possibilities is given in Maling (1960). The variations in how the relationships between maximum and minimum particular scales may be changed give rise to different numerical values for the constants of the projection, and therefore to the location of the standard parallels. This, in turn, creates a considerable number of possibilities in choosing between different conical projections; therefore it is desirable to see what practical guiding principles can help to make a logical choice. In the study of the conical projections

the underlying assumption is made, but not always recognised, that every part of a zone to be mapped has equal importance. In other words we assume that a country completely fills the fan-shaped outline of a conical projection between the limiting parallels and meridians. This assumption is clearly unrealistic if we want to produce a map of Argentina, India, Mexico or Norway on a conical projection, because the countries are asymmetrical, showing much variation of width of land with latitude. Therefore the derivation of projection constants which depend only upon scale ratios between the centres and edges of the map must be misleading. This subject has been studied by Kavraisky, who proposed the use of a constant to help make the choice of suitable standard parallels for conical projections which takes the shape of the country into account. Rewriting equations (10.42) and (10.43) (pp. 208–209) in the form

$$\varphi_2 = \varphi_N - (\varphi_N - \varphi_S)/K \tag{11.01}$$

and

$$\varphi_1 = \varphi_S + (\varphi_N - \varphi_S)/K \tag{11.02}$$

the constant K may be varied according to the shape of the country to be mapped. Kavraisky's values for K may be listed as follows, for the shapes indicated in Fig. 11.08:

Small extent in latitude but large extent in longitude, $K = 7$
Rectangular outline with longer axis north–south, $K = 5$
Circular or elliptical outline, $K = 4$
Square outline, $K = 3$

A more sophisticated approach was used by him to derive the *Conical equidistant projection* (*Kavraisky IV*) originally intended for a map of the European part of the USSR. This made use of a least-squares analysis to obtain the projection constants n and C, using the land area in every 1° belt of latitude as a weighting factor. His method of obtaining the constants has been described in detail by Maling (1960).

A somewhat different example of modification to a minimum-error representation makes use of the establishment of geometrical conditions round the periphery of the region to be mapped. The best-known of these is the *Chebyshev condition*, originally stated as long ago as 1856, which is the statement that a region may be best shown conformally if the sum of the squares of the scale errors over a region is a minimum. Although referring specifically to conformal projections it is, of course, the concept of minimum-error representation already described. Chebyshev further suggested, though this was not proved until later, that this results if the region is bounded by a line of constant scale. This condition is satisfied in the Stereographic projection, which is always bounded by a circle of constant scale. However, later development of the theory made it possible

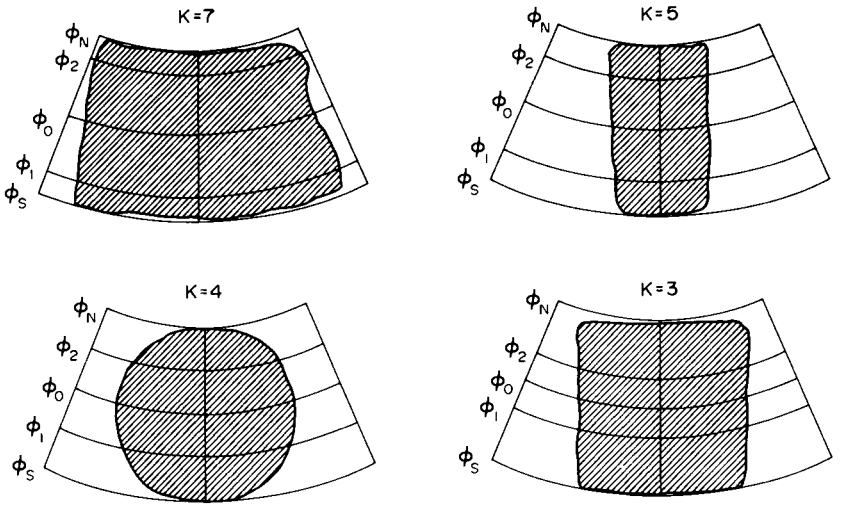


FIG. 11.08 Definition of Kavraisky's constant, K, to aid the choice of standard parallels for conical projections to show countries having different shapes.

to bound the map by other lines of constant scale, notably by ovals. This has led to the description and use of several useful projections for maps of the major continents or oceans. About 1944, Ginzburg had applied much the same approach to the Azimuthal equal-area projection and produced the *TsNIIGAiK projection with oval isolines*, which Maling (1960) called *Ginzburg III*. This projection was used for a map of the Atlantic Ocean which has appeared in the *Atlas Mira* (1945), and several later publications.

In 1953, O. M. Miller applied the Chebyshev conditions to an oblique aspect Stereographic projection to produce the *Miller prolated Stereographic projection*, this having been designed for a minimum-error conformal map of Europe and Africa. He subsequently applied the same technique to produce a similar map for Asia and Australasia, which also has oval isolines. Later applications of the Chebyshev criterion to the Stereographic projection include the description by Lee (1974) of a map for the Pacific Ocean. It has also been used by Snyder (1984, 1987a) for the *GS-50 projection* prepared for the USGS to represent all 50 states of the USA in their correct geographical relationships without creating undue distortion in the vicinity of Hawaii, Alaska or Florida.

Transformation of a projection by the creation of a pole-line

At first sight it may seem that the presence of a singular point on a map projection is inconvenient, for this means either that the map is abruptly

terminated by a line, or that there is no real edge to the map in that part. In the normal aspect Cylindrical equal-area projection the representation of the geographical pole by two lines of length equal to the equator creates a squat rectangular shape which makes it unattractive for use as a world map. In contrast, the normal aspect Sinusoidal projection looks better, because the geographical poles are represented by points and the meridians converge to them.

However, a defect of the Sinusoidal projection, shared also by Mollweide's projection (Fig. 6.07, p. 117) is the large amount of angular deformation towards the edges of the map. On the Sinusoidal projection $\omega > 90^\circ$ and on Mollweide's projection $\omega > 80^\circ$; this deformation is clearly evident from the obliquity of graticule intersections towards the edges of the map in high latitudes. It is easy to imagine that the substitution of a short *pole-line* would reduce angular deformation by making every graticule intersection close to a right angle. This may be done by using constants which create singular points at the geographical poles in the normal aspect or the corresponding points in the other aspects. The length of the pole-line is governed by the choice of constants. A common choice is for it to be one-half the length of the equator. The shape of the line matches the parallels in the normal aspect. Thus the pseudocylindrical projections like the sinusoidal and Mollweide's projections, which all have rectilinear parallels, will also have a straight pole-line. Figure 13.05 illustrates this for the recentred version of the *Eckert VI pseudocylindrical projection*, which is an equal-area projection having sinusoidal meridians. Such a map may be called *truncated* or *flat-polar*. Other classes of projections with curved parallels may be similarly modified to have curved pole-lines. The *Aitoff–Wagner projection*, illustrated in Fig. 1.05, p. 8, shows this. We do not derive the algebraic expressions for this kind of modification in this book, though the coordinates needed to compute certain projections with pole-lines are given in Appendix I, pp. 432–441. The reader who wishes to investigate the general theory of this kind of transformation is referred to Wagner (1949, 1982).

CHAPTER 12

Choosing a suitable map projection – the graphical and analytical methods

> There is much to be said for the belief that the best way of judging a world-projection is to look at it.
>
> A. R. Hinks, *Geographical Journal*, 1934

Introduction

In Chapter 11 we saw that, for most practical purposes, the choice of a projection for a particular map is governed by the need to keep deformation as small as possible, and that some ingenuity may be required to accomplish this in designing a map for a particular country and purpose. An important way of achieving this aim is to choose the origin and aspect of a projection in such a way that the area to be mapped is located in that part of the projection where distortion is least. The graphical and analytical methods to be described in this chapter have largely evolved from this idea.

Graphical methods of selection by visual comparison of overlays

This method allows the choice of class, and often the special property of a projection, by using the patterns of distortion isograms for different projections plotted on transparent plastic and making visual comparisons between them. This is really the only simple way of comparing the relative merits of those classes of projection in which the isograms have more complicated patterns than those for the cylindrical, conical and azimuthal classes. The primary requirement is for the isograms for different projections to be plotted at the same principal scale, e.g. 1/20 000 000 for maps of large countries or continents and about 1/100 000 000 for world maps. There is no need to show the parallels or meridians on these maps; indeed it is less confusing if they are not plotted. However, it is important to indicate the origin and axes to which the isograms are related, and

obviously the lines of zero distortion are also useful. The overlays may be placed singly or in groups over a rough outline sketch-map of the country or continent drawn at the same scale. By shifting the position and orientation of the overlay it is possible to estimate any advantage to be gained from a change in origin or change in orientation of the lines of zero distortion. What we are attempting to achieve by these means is the idea contained in Chapter 7, p. 137, that the patterns of distortion possessed by a given projection remain constant however much we change the aspect of the projection. We are therefore using the overlay as a frame through which we can imagine how the distortion will occur, just as an artist may compose a picture by looking at objects through a small rectangular cardboard frame, or a photographer uses the rectangular ground-glass screen of the camera viewfinder.

When two or more overlays for different projections are superimposed it is easy to compare extreme values for p or ω from the isograms. Figure 12.01 illustrates such a comparison by combining the ω isograms for Bonne's projection and for the Azimuthal equal-area projection which have been plotted to the same principal scale and brought into coincidence for an origin in latitude 45°N, longitude 100°W. This indicates that the extreme values of ω, encountered in Alaska and Greenland are about 5–8° on the Azimuthal equal-area projection but greater than 15° on Bonne's projection. The evident conclusion is that the Azimuthal equal-area projection is to be preferred to Bonne's projection as the base for an equal-area map of the North American continent. The procedure is now repeated with any other equal-area projections which are deemed to be suitable and for which suitable overlays have been prepared. However we have only compared the *maximum* values for ω round the edges of the map. Perhaps it would be more sensible to confine our attention to the centres of each map and compare, by measurement those areas for which $\omega < 1°$ or $\omega < 5°$ on each of the overlays. This approach has been used by Robinson (1952, 1953) to evaluate the suitability of various world map projections, and especially to measure the advantages which different kinds of modification have upon them. We return to this subject in Chapter 13, pp. 275–277.

It must be realised that the underlying map is only a rough guide. If an overlay is to be compared with a map, the relationship between the isograms and the map outline is only precisely true for that aspect and projection upon which the map was compiled. The detailed outlines of the coastline or international boundaries are altered even if the aspect of the projection is only slightly changed and, of course, the outlines vary uniquely for every other projection. Consequently the visual comparison between the map and overlay cannot be exact, and this is why we only recommend and illustrate a rough sketch-map. The purpose of this outline is to indicate approximately the extent of the country or continent. Where

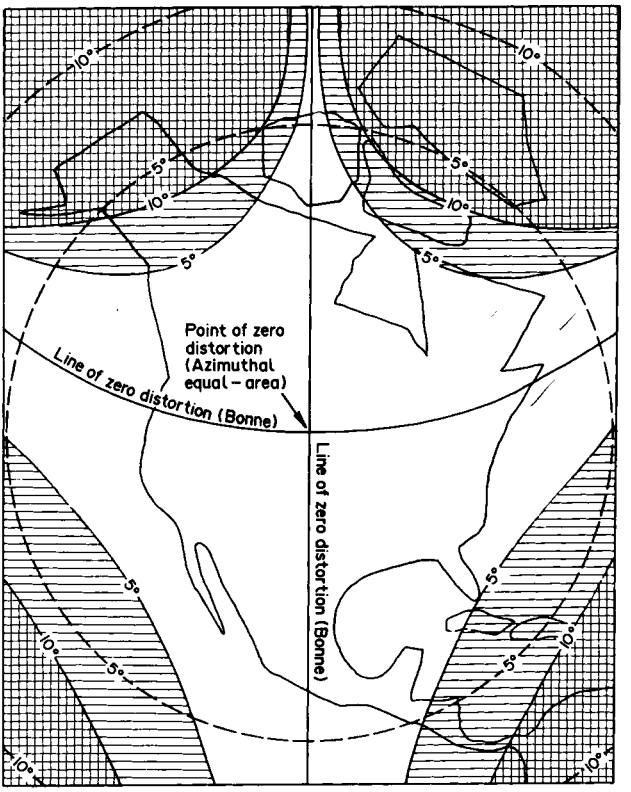


FIG. 12.01 The comparison of the relative merits of Bonne's projection and the Azimuthal equal-area projection for a map of the North American continent. Both of these are equal-area projections so that the best way of comparing them is through maximum angular deformation ω. The origin of both projections is the point with latitude 45°N, 100°W. Isograms for maximum angular deformation are shown for both projections at intervals of $\omega = 5°$ and 10°. The patterns refer to the isograms for Bonne's projection. Note that the coastlines are drawn roughly to indicate their approximate location. They do not coincide with their positions on either of these projections accurately, and are only an approximate guide to the extent of the area to be mapped.

two or more projections are being evaluated, the required comparison is to be made between the distortion isograms. If these have been carefully plotted to the same principal scale, the designer can obtain a fairly accurate impression of the relative merits of different projections.

Figure 12.02 shows another example of comparing two projections; namely a comparison between the Conical equidistant projection with two standard parallels (de l'Isle) and the Azimuthal equidistant projection for a proposed map of China.

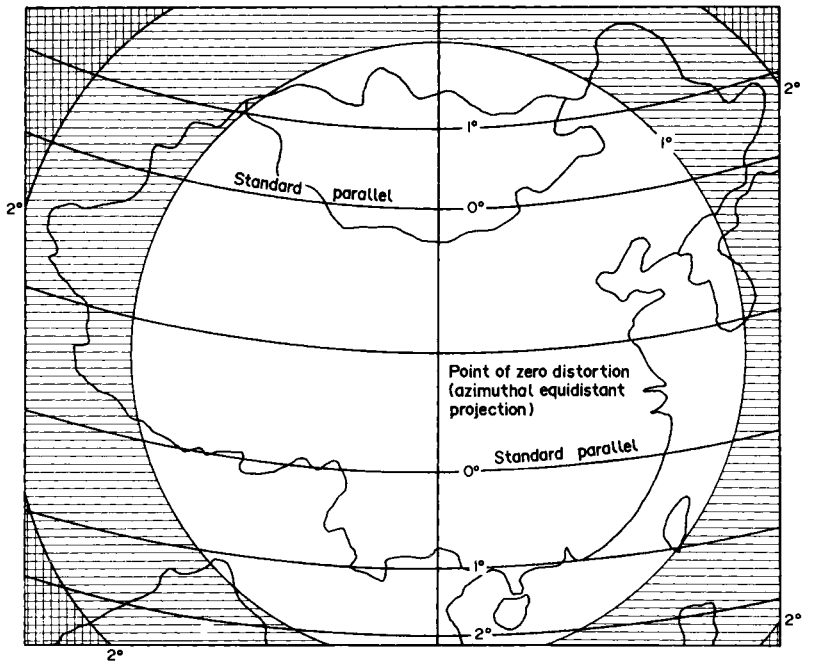


FIG. 12.02 The comparison of the relative merits of the Azimuthal equidistant projection (Postel) and the Conical equidistant projection (de l'Isle) for a map of China. The origin of the Azimuthal equidistant projection is the point in latitude 35°N, longitude 105°E, and the corresponding graticule intersection of the Conical equidistant projection is made to coincide with this. Either area scale (p) or maximum angular deformation (ω) might be compared for these projections. Here the isograms of maximum angular deformation at $\omega = 1°$ and 2° have been plotted. The patterns refer to the isograms for the Azimuthal equidistant projection. Note that the coastlines and frontiers are sketched roughly to indicate their approximate location. They do not coincide with positions in either of these projections accurately, and are only an approximate guide to the extent of the area to be mapped.

The combined analytical and graphical method of selection

Although methods like the use of Kavraisky's constant, K, may be valuable in certain kinds of choice, they only represent a partial solution to the larger problem of deciding if modification of certain projections is going to be helpful in producing a better map. We have to devise a systematic method of investigation, and in seeking this we cannot do better than extend the graphical methods already described and employ the simple analytical techniques briefly described by Miller (1941). In order to show how these may be applied to a specific problem we select the example which Miller himself described, namely to find a conformal projection suitable for a single map of Latin America. This study led

ultimately to the description of the *Bipolar oblique conformal conical projection* (Fig. 11.03, p. 231) which represents the whole of the New World in a single map.

The choice of a conformal map for Hispanic America

In order to proceed with the analytical part of the investigation it is necessary to specify certain limiting values of distortion which we wish to satisfy on the map. For a conformal map we might specify that the area scale should always lie between two limits such as $0{\cdot}95 < p < 1{\cdot}05$, which is equivalent to the statement that distortion of area never exceeds $\pm 5\%$. Alternatively, we might specify, like Miller, that the particular scales should lie within the range $0{\cdot}965 < \mu < 1{\cdot}035$, or, in other words, that linear distortion does not exceed $\pm 3{\cdot}5\%$. We should note that there is nothing magical about the choice of these numerical values for area scale and particular scales. The choice of these is quite arbitrary, but has to be realistic. We would not be able to produce a map for the whole of Latin America if we specified that $0{\cdot}999 < \mu < 1{\cdot}001$. On the other hand, the investigation would not be particularly rewarding if we specified that $0{\cdot}5 < \mu < 2{\cdot}0$, because a large number of projections would satisfy these conditions and the selection between them would not be helped.

The area to be mapped is illustrated in Figs 12.03, 12.04 and 12.05. It represents the whole of the continent of South America and also Central America, extending from the northern frontier of Mexico in latitude 32°N near the Gulf of California. A preliminary study suggests that the origin of the projection might be located at $\varphi_0 = 0°$, $\lambda_0 = 72°\text{W}$. Young's rule gives $z/\delta \approx 1{\cdot}4$, which is so close to the critical value for a conformal projection that it is debatable whether an azimuthal, cylindrical or conical projection is to be preferred. In his study of the subject Miller compared modified versions of the Stereographic projection, normal aspect Mercator's projection and the Transverse Mercator projection before finding a satisfactory solution in the choice of an oblique aspect Conformal Conical projection. We begin by investigating the possible use of the Stereographic and two versions of the Mercator projection without modification for both the methods, and the results are most instructive. We investigate each of the projections in turn to determine the location of the limiting isogram for $\mu = 1{\cdot}035$ and plot the result in Fig. 12.03.

The study of the separate projections may be summarized as follows.

Transverse aspect stereographic projection

From Eppendix I, p. 433, the equation for the particular scale of the Stereographic projection is

$$\mu = a = b = \sec^2 . z/2$$

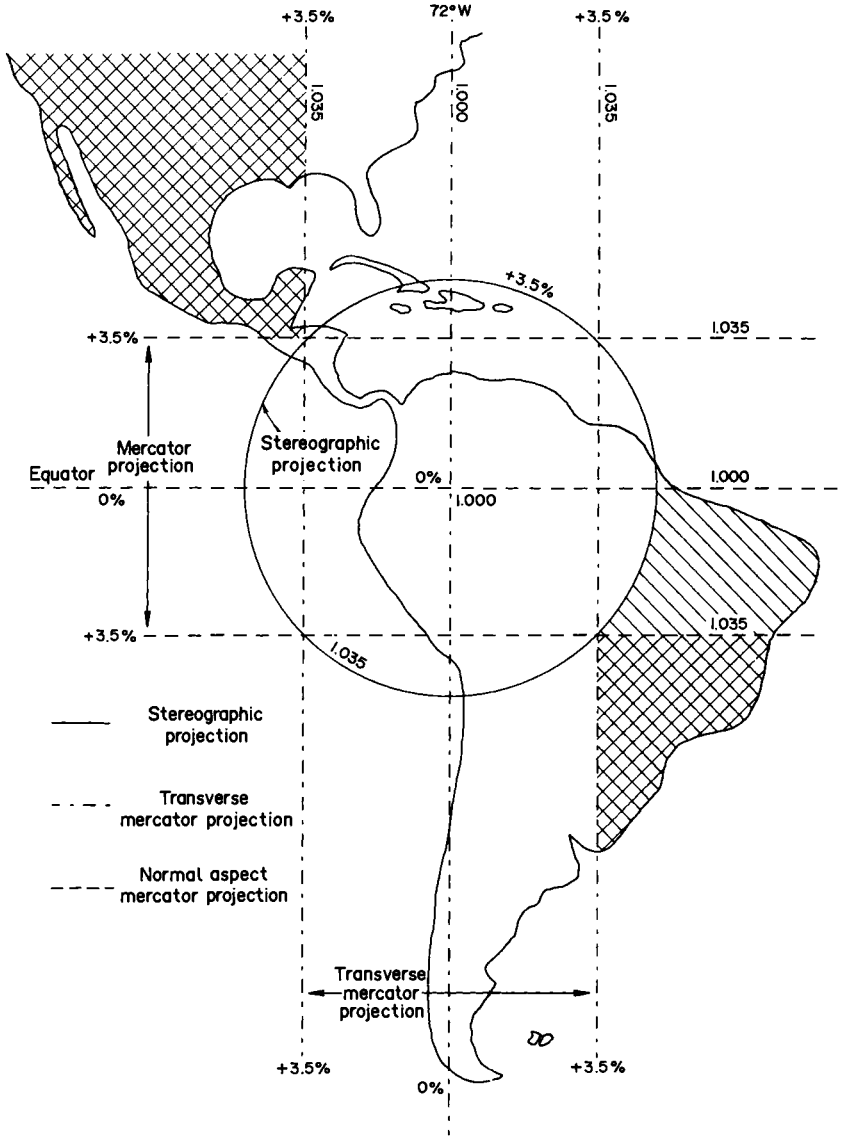


FIG. 12.03 Graphical comparison of the isograms for particular scales on the normal aspect Mercator projection, Transverse Mercator projection and the Stereographic projection for a conformal map of Latin America. The same technique is used as illustrated in Figs 12.04 and 12.05. The origin for the Stereographic projection is the point on the equator in longitude 72°W. The line of zero distortion for the normal aspect Mercator projection is the equator, and that for the Transverse Mercator projection is the meridian 72°W. Thus figure compares the regions enclosed by the isograms for particular scale $\mu = 1{\cdot}035 = +3{\cdot}5\%$. To aid interpretation the parts of the region where this particular scale is exceeded on the Stereographic and Transverse Mercator projections are shaded.

Thus if we employ a transverse aspect stereographic projection, there is a single point of zero distortion at the origin, on the equator in longitude 72°W. Here the principal scale is equal to unity and the particular scale increases radially outwards to the specified limit (+3·5%) where

$$\sec^2 z/2 = 1{\cdot}035$$

Solving this equation we find that $z = 21°12'$, so that the only part of Latin America which can be mapped to the required specification lies with the circle shown in Fig. 12.03.

Modified transverse stereographic projection

We specify that the lower value for the particular scale, $\mu = 0{\cdot}965 = A$ is preserved at the origin of the projection. Then the point of zero distortion is replaced by a standard circle of angular distance z_a from the origin. It can be shown that

$$\cos^2 z_a/2 = A \qquad (12.01)$$

and the formula needed to define the modified projection for a sphere of unit radius may be written

$$r = 2A \,.\, \tan z/2 \qquad (12.02)$$

with the equation for the particular scales

$$\mu = A \,.\, \sec^2 z/2 \qquad (12.03)$$

Putting $\mu = A = 0{\cdot}965$ at the origin of the projection, we obtain the radius of the standard circle to be $z_a = 21°34'$. Substituting $\mu = 1{\cdot}035$ and $A = 0{\cdot}965$ in equation (12.03) the upper limit of particular scale is found from the distance $z = 30°08'$ from the origin. Figure 12.04 shows the position of the standard circle and the isogram for $1{\cdot}035 = +3{\cdot}5\%$. Comparison of Figs 12.03 and 12.04 indicates the value of introducing this kind of modification to an azimuthal projection.

Normal aspect Mercator's projection

From the study of this projection in Chapter 10 we know that the line of zero distortion is the equator and that the particular scale is located along the parallels where $\sec \varphi = 1{\cdot}035$ or

$$\varphi = \pm 14°57'$$

The part of Latin America which can be mapped according to this specification is shown in Fig. 12.03.

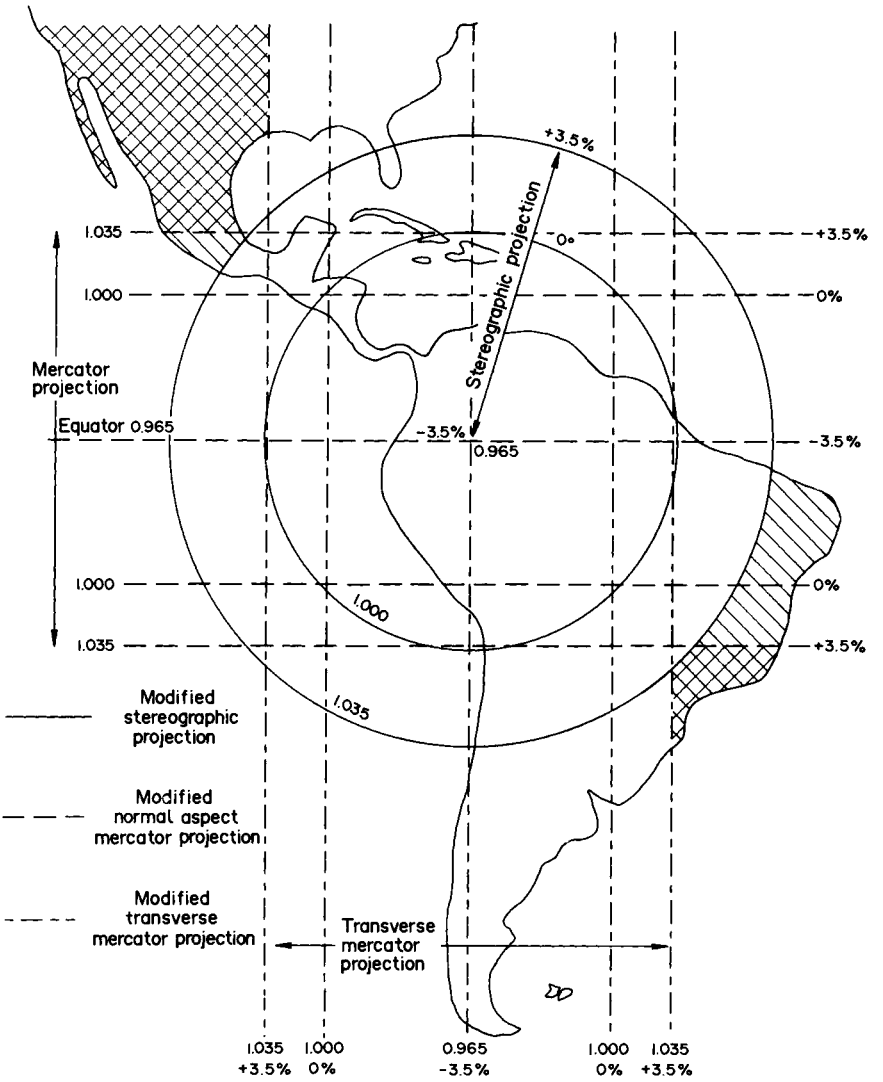


FIG. 12.04 Graphical comparison of the isograms for particular scales on modified versions of the normal aspect Mercator projection, Transverse Mercator projection and Stereographic projection for a conformal map of Latin America. In this figure the range of particular scales is made $0{\cdot}965 < \mu < 1{\cdot}035$. Therefore the particular scale $\mu = 0{\cdot}965$ ($-3{\cdot}5\%$) is preserved at the origin of the Stereographic projection, along the equator for the normal aspect Mercator projection and along the meridian 72°W for the Transverse Mercator projection. To aid interpretation the parts of the region where the particular scale $\mu > 1{\cdot}035$ or $+3{\cdot}5\%$ on the Stereographic and Transverse Mercator projections are shaded. Comparison of Figs 12.03 and 12.04 indicate how modification of any of these projections extends the size of the region within which they may be usefully employed.

Modified normal aspect Mercator's projection

If we specify that the limiting particular scale $\mu = 0{\cdot}965$ is preserved at the equator, this has the effect of introducing two lines of zero distortion forming two standard parallels. From equation (10.71), p. 213, we have, for the particular scale at the equator

$$0{\cdot}965 = \cos \varphi_0/1$$

or

$$\varphi_0 = \pm 15°12'$$

Substituting this value in equation (10.71) we now obtain the latitude, φ, where the particular scale becomes the upper limiting value. Thus

$$1{\cdot}035 = 0{\cdot}965/\cos \varphi$$

$$= \pm 21°12'$$

The part of Latin America which can be mapped according to this specification is shown in Fig. 12.04.

Transverse Mercator projection

Since a change in aspect does not alter the position or pattern of distortion isograms, we use the reasoning already used for the normal aspect Mercator to obtain the corresponding numerical values for the Transverse Mercator projection. The only difference is the orientation of the isograms which are now perpendicular to those shown in the normal aspect. In the first case, illustrated by Fig. 12.03, the line of zero distortion is the central meridian lying in longitude 72°W and the limiting scale $\mu = 1{\cdot}035$ occurs at an angular distance $z = 14°57'$ from it. At the equator, therefore, this particular scale occurs in longitudes 57°03′W and 86°57′W respectively. Since the distortion isograms are coincident with small circle arcs parallel to the central meridian, they do not coincide with these meridians elsewhere.

Modifications to the Transverse Mercator projection so that the particular scale $\mu = 0{\cdot}995$ is preserved along the central meridian creates two lines of zero distortion at $z = 15°12'$ on either side of the origin (intersecting the equator at longitudes 56°48′W and 93°12′W on the equator respectively). In Chapter 16 we will find that a similar kind of modification is commonly applied to the Transverse Mercator projection in the form used for topographical maps.

The advantages of modification are evident from the comparison of Figs 12.03 and 12.04 because each of the modified projections which have been studied show that a much larger part of Latin America can be shown within the specified limits of particular scale. In both illustrations we can

TABLE 12.01 *Percentage scale distortion for three map projections to show Latin America based upon calculations by O. M. Miller (1941)*

	Average	Maximum
Modified Stereographic	5·6	18·4
Modified Transverse Mercator	6·3	27·7
Modified normal aspect Mercator	10·8	58·4

see that there is not much to choose between the Stereographic and the Transverse Mercator projections, and that both of these are clearly superior to the normal aspect of Mercator's projection. However the visual impression is obtained from the study of the positions occupied by the isograms for a single value of μ, and does not take into account the magnitude of the particular scales beyond the specified limit. Miller calculated these for a network of 49 points within the area to be mapped (at intervals of 8° latitude and 12 ° longitude) and found that the average and maximum scale distortions for the three projections studied were as shown in Table 12.01.

These figures suggest the modified stereographic projection is the best choice *from these three*, but as the visual appraisal showed, it was closely followed by the Transverse Mercator projection. Nevertheless, all three of them significantly fail to meet the initial specification that linear distortion should everywhere lie between $+3{\cdot}5\%$ and $-3{\cdot}5\%$. It can be argued that this specification cannot be met and therefore somewhat lower tolerances should be tried, for example to find a conformal projection in which the linear distortion does not exceed $\pm 4\%$ or $\pm 5\%$. If this expedient is adopted, the analysis must be repeated to calculate and plot new positions for the limiting isograms. The reader is invited to substitute the conditions that $0{\cdot}95 < \mu < 1{\cdot}05$ and obtain values corresponding to those derived on the last three or four pages.

The oblique aspect Conformal Conical projection

However, relaxation of the specification is justified only when it is clear that no other projection will meet the original specification. After all, we have tried only three possibilities, and have not yet considered use of a Conformal Conical projection. This, in fact, provides a solution which is superior to the other examined above.

From the study of a globe, which is always invaluable for the initial stages of such an investigation, Miller and Briesemeister found that a small circle can be drawn about a pole located in the southern Pacific Ocean, the arc of which divides Latin America into two parts which are

roughly equal in area. By trial and error they found that a pole located in latitude 20°S, longitude 110°W and a small circle 52° distant from it meets this requirement. The problem now is to define the Conformal Conical projection in which this arc forms the middle of the map. Since the concept of an oblique aspect conical projection is unfamiliar, we have, in Fig. 12.05, treated the problem as if we were dealing with a normal aspect conical projection in which the various small circles would be parallels of latitude. Thus we must define a sequence of concentric circular arcs which we have labelled as follows:

z_0 is the small circle with angular distance 52° from the pole which passes through the middle of Latin America. On a normal aspect map this would be φ_0.

z_4 is the small circle passing through the further limit of the map, i.e.

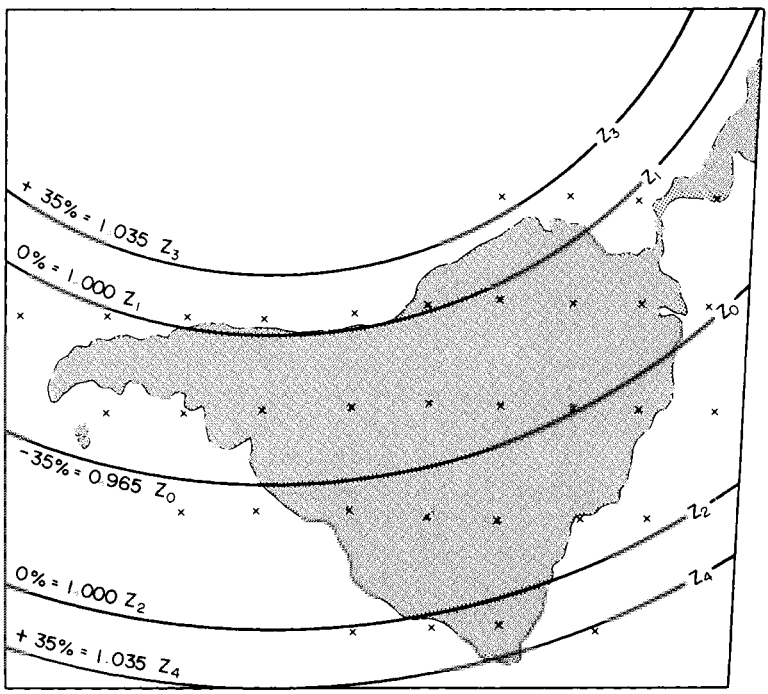


FIG. 12.05 Diagrammatic representation of the isograms for the oblique aspect Conformal Conical projection selected by Miller for the American Geographical Society map of Hispanic America. The isograms are for particular scales $\mu = 0{\cdot}965$ ($-3{\cdot}5\%$), $1{\cdot}000$ (0%) and $1{\cdot}035$ ($+3{\cdot}5\%$), and indicate the extent of the region which can be mapped without exceeding linear deformation $\pm 3{\cdot}5\%$. The crosses indicate those graticule intersections for which Miller calculated the particular scales in his statistical analysis of the merits of this projection compared with the versions of the Stereographic and Mercator's projection already studied.

lying in the Atlantic Ocean and close to the coast of Brazil. On a normal aspect map this would be φ_s.

z_3 is the small circle passing through the nearer limit of the map, in the Pacific Ocean near the Galapagos Islands. On a normal aspect map this would be φ_s.

z_1 and z_2 are two small circles lying between the centre and edges of the map, equidistant from z_0 which will serve as the standard parallels. On a normal aspect conical projection these would be φ_1 and φ_2 respectively.

Compare Fig. 12.05 with Fig. 11.03, p. 231, to see these changes in notation.

Inspection of a map or globe shows that where the pole is in latitude 20°S, 110°W, $z_0 = 52°$, $z_4 = 73°$ and $z_3 = 31°$. The shape of the area to be mapped suggests that the use of Kavraisky's constant $K = 7$ may be a suitable choice for the location of the standard parallels. By trial and error, Miller found $z_2 = 36°20'$ and $z_1 = 66°35'$ would best meet the specification. This corresponds to the use of $K \approx 8$. The resulting map has the lower limit of particular scale $\mu = 0{\cdot}965$ along the centre with the upper limit $\mu = 1{\cdot}035$ along the small circle arcs z_4 and z_3.

Analysis of the particular scales at the 49 points on the map gave an average percentage scale distortion of $\pm 1{\cdot}8\%$ rising to a maximum value of 9·8%. Comparison of these figures with those in Table 12.01 indicates immediately that this projection is greatly superior to any of the three studied earlier. However, the really convincing display is through graphical representation of the limiting isograms which are shown in Fig. 11.03, p. 231.

The next logical step was for Miller to locate a second pole in the North Atlantic Ocean suitable for mapping the remainder of the North American continent on a second oblique conformal conical projection. By combining the two we have the *bipolar* version illustrated on p. 231. The two parts of the map join along a straight line running through the Caribbean and Central America. However, the correspondence between the two projections is not exact along this line and a small amount of fudging had to be applied in order to create the appearance of a continuous map.

The choice of a suitable projection for CORINE

We turn now from a classic investigation for a small-scale sheet or atlas map to the application of similar methods to choose a projection to be used internally within a geographical information system. The example described here is that for CORINE, this being the acronym for *Coordinated Information on the European Environment*. The nature of this GIS and the initial uses proposed for it have been described by Wiggins *et al.*

(1986) and many earlier papers. Briggs and Mounsey (1989) is the most up-to-date statement at the time of writing. The following investigation was undertaken by the author at the request of Dr Mounsey in the spring of 1989. At that time the databases contained information on topography, climate, water resources, sites of scientific interest (biotypes) and soils. Other variables were to be added as and when reliable information became available.

The database is being constructed through the Geographical Information System known as ARC/INFO. This is a well-known software system in which the processing (by ARC) of cartographic data in vector mode is coupled with the processing (by INFO) of the associated attribute data. This software was written by ESRI (Environmental Systems Research Institute) in the middle 1980s. The data which have been collected are stored as a series of layers in the database. If satisfactory comparison between each layer is to be attempted, for example to compare vegetation with soil type, all these data must obviously be stored on the same projection. Although ARC/INFO allows conversion between 20 different projections, the choice of the projection to be selected for storage of the data is critical, if only to save computer time and storage space which would be wasted if the initial data all had to be treated separately each time it has been accessed. Chapter 11 included some comments by Briggs and Mounsey (1989) about the difficulty of fitting the soil map of the EC to the topographical base. It was for this reason that the present author was asked to look at the general problem of 'What is the best projection for the European Community to be used in the CORINE GIS?'

In making the following analysis the author used virtually the same technique as that outlines for the Hispanic America map, saving only that, because the area to be mapped is much smaller, the work was more conveniently done on a map rather than a globe. Most of the conclusions may be drawn from a good atlas map of Europe, such as that at scale 1/12 900 000 forming Plate 49 of *The Times Atlas of the World*, 1967 edition, which is on Bonne's projection.

The member countries of the European Community straggle obliquely across the map of Europe. The lack of compactness is characterized by two major linear trends. The most obvious of these is the arcuate alignment of countries along a curve from Scotland through Britain, France and Italy into the eastern Mediterranean. We shall call this the *Glasgow–Lyon–Alexandria arc*. It is the arc of a small circle of spherical distance $z = 25°$ from a pole situated in latitude 55°N, longitude 45°E in northern Russia. The second is the great circle arc which appears on the Lambert conformal conical projection of Europe as the straight line passing through Cape St Vincent–Frankfurt–Gdansk, which we will call the *St Vincent–Gdansk axis*. Evidently the two lines intersect close to Lyon. The

geometrical centre of the EC appears to lie further east, in Bavaria, close to the small town of Tuttlingen (latitude 48°N, longitude 9°E).

Albers' projection

We naturally think of a conical projection as being the most suitable map for Europe, if only because this is what we find in most atlases, and because Hinks said that this must be so. It also seems that for most purposes an equal-area projection would be preferable. Consequently the obvious choice for an equal-area map of Europe is *Albers' projection.* This had already been chosen as a good estimate of what might be needed, and the author had to compare other possible projections against the version in which the standard parallels are located in latitudes $\varphi_1 = 44°\text{N}$ and $\varphi_2 = 53°\text{N}$, illustrated in Fig. 12.06. It follows from the equations for Albers' projection in Appendix I that the central parallel is $\varphi_0 = 48°{\cdot}5\text{N}$ and $\delta = 4°{\cdot}5$. Then the constant of the cone is $n = 0{\cdot}7466469$ and the integration constant $C = 1{\cdot}5547788$. The particular scales and distortion characteristics are given in Table 12.02.

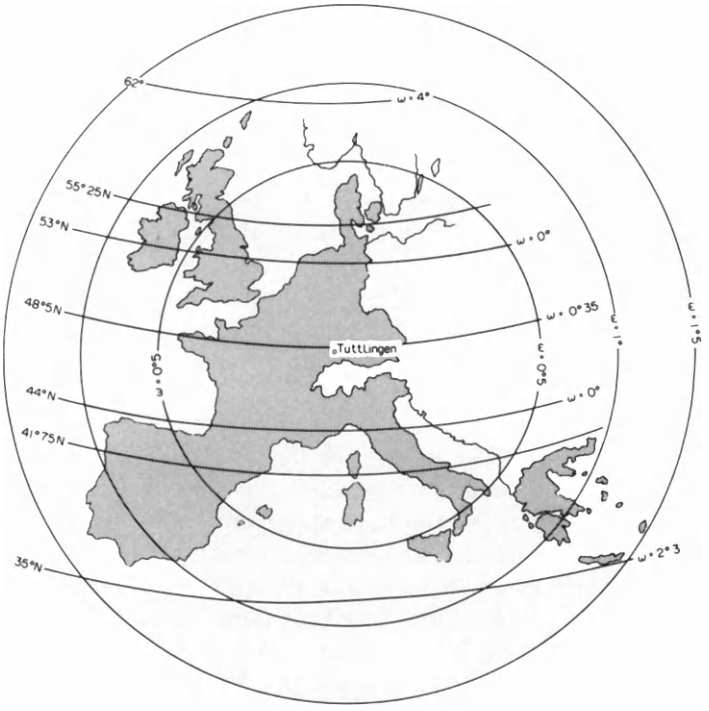


FIG. 12.06 The location of the member countries of the European community plotted on Albers' projection (No. 14a in Appendix I) and showing the values for ω on certain parallels, which are also those in Table 12.02 for a projection with standard parallels in latitudes 44°N and 53°N. Superimposed upon this map are the circular isograms for an Azimuthal equal-area projection centred on Tuttlingen.

TABLE 12.02 *Particular scales and distortion characteristics for Albers' projection with standard parallels $\varphi_1 = 44°N$ and $\varphi_2 = 53°N$ for mapping the European Community*

Latitude φ	Particular scales: Meridian h	Particular scales: Parallel k	Percentage scale: error $(a-1=1-b)\%$	Area scale p	Maximum angular deformation ω
62	0·9568	1·0353	3·5	1·0000	3°·98
53	1·0000	1·0000	0·0	1·0000	0°
48·5	1·0031	0·9969	0·3	1·0000	0°·35
44	1·0000	1·0000	0·0	1·0000	0°
35	0·9803	1·0201	2·0	1·0000	2°·28

Strictly speaking, however, the positions of the standard parallels in Albers' projection should be midway between the middle and limiting parallels of the map. Since $\varphi_4 = 62°N$ and $\varphi_3 = 35°N$ are specified by the limits of the territory and the central parallel is, as before, 48°·5N, the standard parallels ought to be in latitudes $\varphi_1 = 41°·75$ and $\varphi_2 = 55°·25N$. It follows that $n = 0·7437643$ and $C = 1·5471196$. The new set of particular scales are given in Table 12.03. This indicates a small improvement in the map.

Murdoch's third projection

Despite the fact that *Murdoch's third projection* is not rigorously equal-area, it does have the special property that the total area mapped is represented without distortion, and it is the minimum-error conical projection. Therefore it is interesting to compare the results for this projection with Albers' when specified for the same bounding parallels. The constant of the cone is $n = 0·762457$ and the integration constant, $C = 0·851917$.

TABLE 12.03 *Particular scales and distortion characteristics for Albers' projection with standard parallels $\varphi_1 = 41°·75N$ and $\varphi_2 = 55°·25N$ for mapping the European Economic Community*

Latitude φ	Particular scales: Meridian h	Particular scales: Parallel k	Percentage scale: error $(a-1=1-b)\%$	Area scale p	Maximum angular deformation ω
62	0·9711	1·0297	3·0	1·0000	3°·36
55·25	1·0000	1·0000	0·0	1·0000	0°
53	1·0042	0·9958	0·4	1·0000	0°·5
48·5	1·0069	0·9931	0·6	1·0000	0°·8
44	1·0035	0·9964	0·3	1·0000	0°·4
41·75	1·0000	1·0000	0·0	1·0000	0°
35	0·9834	1·0169	1·7	1·0000	1°·92

TABLE 12.04 *Particular scales and distortion characteristics for Murdoch's third projection for the latitude belt 35°N to 63°N. Standard parallels and* $\varphi_1 = 40°{\cdot}7$ *and* $\varphi_2 = 58°{\cdot}7$

Latitude φ	Particular scales: Meridian h	Particular scales: Parallel k	Area scale p	Percentage area scale error $(1-p)\%$	Maximum angular deformation ω
62	1·0000	1·0151	1·0151	1·51	0°·86
60	1·0000	1·0063	1·0063	0·63	0°·36
58·7	1·0000	1.0000	1.0000	0·0	0°
55	1·0000	0·9933	0·9933	0·67	0°·39
50	1·0000	0·9898	0·9898	1·02	0°·59
45	1·0000	0·9939	0·9939	0·61	0°·35
41	1·0000	1·0017	1·0017	0·17	0°·10
40·7	1·0000	1.0000	1.0000	0·00	0°
35	1·0000	1·0204	1·0204	2·04	1°·16

The extreme parallels have particular scales, $k = 1{\cdot}0204$ in latitudes 63° and 35°. This is $\pm 2\%$ scale error. In the middle of the map ($\varphi = 49°$) $k = 0{\cdot}99008$ so that the percentage scale error is barely 1%.

The particular scales are given in Table 12.04, which shows that the area scale varies within the limits $\pm 0{\cdot}02$ and the maximum angular deformation through $\pm 0°{\cdot}6$ except at the extreme edges of the map.

Oblique aspect Albers' projection

The lack of compactness of the countries to be mapped has already been noted. If we sketch the two axes already described on a globe, and then compare the resulting small circle arc with the parallels, we find the best correspondence seems to be with a parallel in approximately 60–65°.

These two axes allow us to imagine the use of an oblique aspect Albers' projection to depict the EC. In order to proceed without numerical analysis in the present example, the author plotted a straight line AB and, constructed two concentric circular arcs with centre A to represent the parallels 60° and 65° on a normal aspect Albers' graticule. Placing the straight line on the St Vincent–Gdansk axis, and trying to fit the circular arc in the position Glasgow–Alexandria, it is easy to see that the smaller radius curve (i.e. of spherical distance or 'colatitude' 25°) fits the alignment of EEC countries better than the 30° curve. The point A is located in northern Russia (latitude 55°N, longitude 43°E) and the two axes intersect near Lyon. The extremes of the map are now Bornholm, which is 15°·9 from A and Cape St Vincent, which is 39°·2 distant. Using these as the equivalent of the limits of the map we find, first, for the Lambert conical equal-area projection with only one standard parallel gives scale errors of 1·4% and 3·6% at these limits (which, it may be recalled,

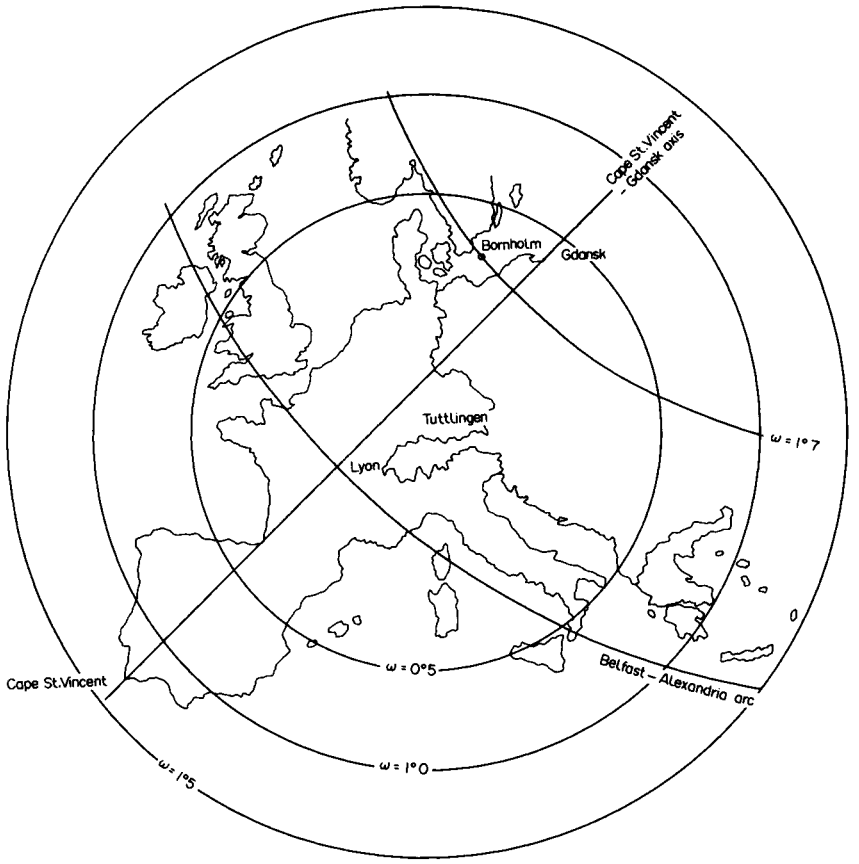


FIG. 12.07 The distortion isograms for an oblique Albers' projection intended to serve as the base for the CORINE GIS, compared with the isograms for an Azimuthal equal-area projection centred on Tuttlingen.

are similar in magnitude to those found for the normal aspect Albers' projection. The oblique Albers' solution provides the results in Table 12.05. This indicates a significant improvement upon the corresponding results for the normal aspect Albers' projection.

TABLE 12.05 *Oblique Albers' projection. Particular scales and distortion characteristics for the limiting extremities which, for this aspect are the north-east and south-west extremes of the map*

	Particular scales		Maximum angular deformation	Percentage scale error
	h	*k*		
Bornholm	0·9851	1·0152	1°·7	1·5
C. St Vincent	0·9843	1·0159	1°·6	1·6

TABLE 12.06 *Azimuthal equal-area projection, particular scales and distortion characteristics for extreme points of the EC measured from the origin in latitude 48°N, longitude 9°E*

	Radial distance z	Particular scales		Percentage scale error $(1-h)\%$	Maximum angular deformation ω
		h	k		
Bornholm	7°·9	0·9976	1·0024	0·24	0°·3
Blasket Islands	13°·1	0·9935	1·0066	0·66	0°·7
Cape St Vincent	17°·2	0·9888	1·0113	1·13	1°·3
Faeroe Islands	17°·1	0·9888	1·0113	1·13	1°·3
Crete	18°·1	0·9875	1·0126	1·26	1°·4
Rhodes	18°·4	0·9871	1·0131	1·31	1°·5

Azimuthal equal-area projection

We see that the extreme distortion values are of similar order to those found with the Azimuthal equal-area projection, which is marginally better because the values for the percentage scale error and ω are slightly smaller. Consequently there is nothing to be gained from using this oblique aspect Albers' projection, though it is an improvement on the normal aspect Albers' considered first.

A simple but useful way of appraising the location of the origin of an azimuthal projection is to plot a series of concentric circles of radii z at the scale of a convenient atlas map and to shift this overlay about on the map until a good fit is obtained between some of the extreme points of the area to be mapped. Using some of those places listed above, the $z = 15°$ circle can be moved about until the centre is more or less equidistant from Rhodes, the Faeroes and Cape St Vincent. This indicates the approximate origin for the projection in latitude 48°N, longitude 9°E already mentioned.

Using this point as the origin, the particular scales for a series of points lying at the extremities of the EC are those listed in Table 12.06. From this table we see that the distortion at the extremities is smaller than may be obtained with either version of Albers' projection or using Murdoch's third projection. Indeed, we have succeeded in keeping maximum angular deformation to 1°·5 or less throughout the whole of the EC.

Towards an automatic method of choice

If the processing of the layers of a GIS is to be undertaken automatically, thereby dispensing with the transparent overlays carefully drawn by the surveyor or planner, presumably there is also an argument that our use of transparent overlays to choose between different projections is equally obsolescent, and should be replaced by an automated system of selection.

There is, of course, no particular difficulty in comparing the distortions

of two or more projections in the manner adopted by Miller, and presented in Table 12.01. We simply write a short program to calculate h and k within nested loops of the required number of meridians and parallels which steps through a sufficient number of graticule intersections. For a comparatively simple example like those class D projections in which the distortion isograms coincide with the normal aspect graticule, it is sufficient to obtain the 60 graticule intersections for a 5° graticule. Polyconic or pseudocylindrical projections having more complicated patterns of isograms might require a denser pattern of points. However, reduction of the mesh size to 2° or even 1° increases the number of graticule intersections to 375 or even 1500 points, and this may add significantly to the amount of computing. The most useful parameter to determine is probably the average percentage scale error, as given in Table 12.01. The reader is warned against attempting any more sophisticated statistical tricks using such data. Because the graticule is a regular systematic pattern, we are, in effect, carrying out two-dimensional systematic sampling, and the absence of any random element in the choice of points means that the majority of statistical techniques, which are based upon the probability of the occurrence of a random event, simply do not apply. See Maling (1989) for additional comments on this subject. However, the method is somewhat clumsy and depends upon the knowledge and experience of the user who wishes to choose a suitable projection. In order to automate choice it is desirable to exercise some kind of disciplined control over the method and logic of choice.

The author is aware of only two attempts to achieve this, neither of which is wholly successful, but which indicate useful directions for future research in this field. The first is the interesting preliminary paper by Bugaevskii (1982); the second is that of Jankowski and Nyerges (1989).

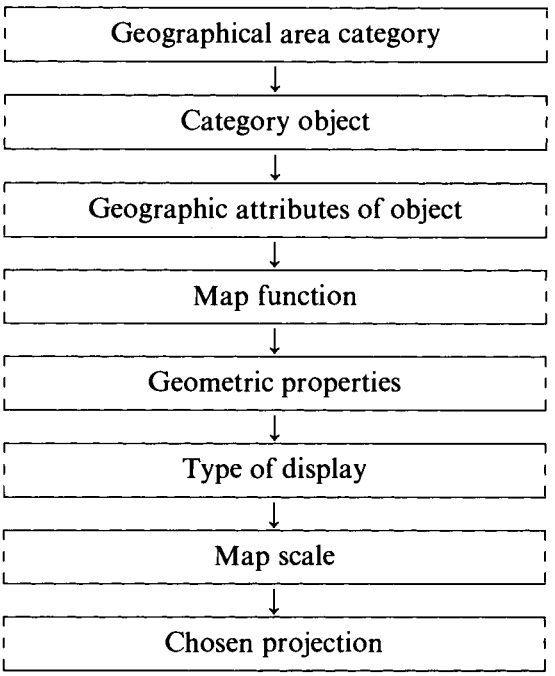
Bugaevsky has attempted to express choice through a rather complicated equation comprising 10 terms each representing a different factor or property of the map projection. Each term derives the sum of squares of a variable, which is suitably weighted to allow for changes in emphasis, corresponding to recognition of the purpose of the map. The individual terms include those which relate the combinations of the particular scales a and b to the distortions which we know already, and in combinations which represent linear, area and angular errors. There are also additional terms which consider variety of expressions each of which must be assigned an unique weight. These include the amount by which a rhumbline and a great circle depart from straight lines.

However, the result is likely to be strongly influenced by what numerical values are assigned as weights, and Bugeavskii does not offer any objective rules to control the choice of these. At present, this appears to be a major difficulty in development of the equation to serve as a basis for automation of choice.

The recent work by Jankowski and Nyerges attempts to tackle the problem of choice through the medium of existing software packages resulting in the series of programs which they have called *MaPKBS* or *Map Projection Knowledge-Based System*. This is evidently still a prototype expert system, and is not the ultimate solution to the problem of computer-assisted map projection solution. It has been implemented in a PC-based expert system shell called *Intelligence/Compiler*.

Much of the preliminary work is involved in deciding how to describe each type of request in qualitative terms. Thus all requests for projections relate initially to geographical areas which are either *ambiguous* or *unambiguous*. To the first category belong the requests of world, hemisphere and regional maps; to the second are those of continents, oceans, seas and countries. The second category is further extended to create *frames* which described the geographical limits more fully. For example the frame *Germany* indicates that this is a country, that it has a north–south directional extent and that its geographical location is mid-latitude. Such frames are stored permanently in the knowledge base.

The data hierarchy which governs selection (known as a *top-down hierarchy* in computer jargon) follows this order, which also corresponds to the logic of the manual methods outlined above. The main headings are:



In this top-down hierarchy the most general concept for the projection selection process, which is the geographical area category, is at the top. Thus at the start of the consultation session the system tries to derive a value for the domain of the map by asking the user: 'What is a category of geographic area for which you would like to find a map projection? Is it world, hemisphere, continent, ocean, sea, country, region or other?'

The degree of concept specialization increases as we move down through the list. If the concept matches the data it becomes established in the solution process. Thus the user is asked about preferences regarding shape, distance preservation, type of display and preferred scale of a map. If conformality is not wanted then equivalence is assumed. In other words there seems to be, at present, no way of obtaining any other projections which do not possess these special properties. The established concept invokes a more specialized concept and this continues until finally a specific map projection is reached.

At the stage when Jankowski and Nyerges published their paper, work on MaPKBS was still in progress. Thus the solution described here cannot be regarded as their final word. Several avenues for system expansion are being considered, such as providing the system with a mechanism for checking the correctness of user-derived geographic parameters such as location and directional extent. Interfacing MaPKBS to a graphics environment is an obvious logical extension.

CHAPTER 13

Discontinuities and deliberate distortions

> The aim of this net is to obtain a map showing all the important habitable lands of the earth, with a true representation of area and a minimum distortion of shape.
>
> C. B. Fawcett, *Geographical Journal*, 1949

Introduction

The method of modification through redistribution of the particular scales is useful when applied to maps for a continent such as Europe or to a large country such as the USA or China, but it is a less satisfactory method of reducing distortion on or near the edges of world maps. This is because the singular points are usually located here. There is not much that can be done to reduce distortion close to the geographical poles if these points are portrayed by lines. We must therefore consider the second possibility of using map projections which have deliberately introduced gaps or other discontinuities within them. To use the concept first introduced on pp. 84–85, distortion on these projections is by stretching rather than tearing.

If we are willing to accept the presence of gaps in the continuity of the mapped surface, it may be possible to reduce the excessive distortions which might otherwise appear. We cannot eliminate them, but at least they may be kept within bounds. There are two different ways in which the excessive distortions near the edges of a world map may be reduced:

- by combining two or more projections to create a *composite projection*;
- by using the technique known as *interrupting* or *recentring* a map projection.

Composite projections

The term *composite map projection* appears to have been first used by Fawcett (1949), although the method is much older. It comprises the juxtaposition of two map projections along a particular line, such as

a parallel of latitude in the normal aspect. Another word, *Dinomic*, has recently been used by Baker (1986), who seems to have been unaware of any earlier work. The technique was also exploited by Goode (1925), whose Homolosine projection was formed by joining the normal aspects of the Sinusoidal and Mollweide (Homalographic) projections along the parallel (40°44′11″·8) which is the same length on both projections. Similarly, Kavraisky combined the normal aspect Mercator projection with the Plate-Carrée projection north of 60°N in order to reduce the particular and area scales in high latitudes. The idea has been used in the design of star-shaped and similar projections where discontinuities arise along the equator, and are caused by the need to change the direction of convergence of the meridians and of the curvature of the parallels at the boundary between the northern and southern hemispheres. Some composite maps are extremely complicated combinations of many contiguous projections. For example, Watts (1970a,b) devised a series of projections containing both pseudocylindrical and pseudoconical elements which map the world in five lobes, using further breaks which originate at 90° of longitude from the central meridian and give rise to two lines depicting the equator. Depuydt (1983) also described a world map based upon a similar scheme which comprises five parts (hence the name *Quintuple projection* to describe it). However, the prize for sheer complexity in design of an elaborate composite projection must be awarded to Macdonald (1968) for his *Optimum Continental projection* of the Old World, which appears to have about 24 separate components. This map contains elements of the Conical equidistant projection (Ptolemy) with one standard parallel, the Polyconic projection and Bonne's projection which are fitted along certain parallels and meridians to create an ingenious jigsaw puzzle.

Recentred or interrupted projections

Recentred map projections are those in which the whole map is derived from a series of contiguous gores, facets or lobes, all of which are based upon the same projection. We have already seen in Fig. 1.07, p. 15 that there is some merit in showing the world by means of two separate maps, either as a pair to show one hemisphere each, or, as shown by Fig. 1.07, by two maps showing more than one hemisphere with a small overlap. There is also value in transforming other projections of the world by introducing artificial interruptions as shown in Fig. 13.05, p. 276 or Fig. 15.01, p. 314. In each example excessive angular deformation or exaggeration of area is checked by using only part of the world map which is repeated, because each portion *is based upon a different point of origin*. Hence we justify the value of the word *recentred* to describe such maps. The alternative word *interrupted* relates only to the appearance of the map.

Dahlberg (1962) has made a scholarly historical study of the methods used, and has made the twofold classification by symmetry or lack thereof:

A: symmetrically interrupted arrangements
B: asymmetrically interrupted arrangements

Although Dahlberg recognised six subgroups of the symmetrical class, as illustrated in Fig. 13.01, three of them are effectively the same procedure applied to different aspects of the cylindrical and pseudocylindrical projections in which continuity is maintained along the great circle which is

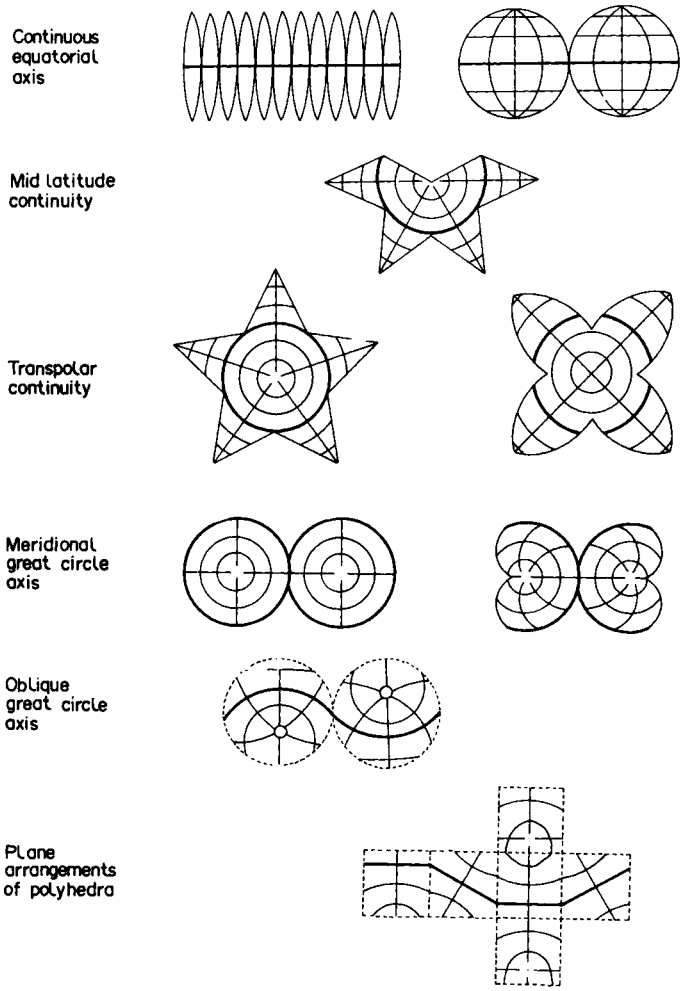


FIG. 13.01 Dahlberg's classification of interrupted map projections. (Source: Dahlberg, 1962.)

the line of zero distortion. Therefore the present author would reduce the number of subdivisions to only three, corresponding to the three ways in which the principal scale is preserved on the projections of Group D. Thus we may regard these as being:

Maling	*Dahlberg*
Continuity along a great circle	Continuous equatorial axis Meridional great circle axis Oblique great circle axis
Continuity along a small circle	Mid-latitude continuity
Point symmetry	Transpolar continuity

We will attempt to use a similar kind of classification to the asymmetrically interrupted projections. However, any neat classification system tends to be obscured by the fact that many of the asymmetrical examples are also composite projections.

An interesting feature of the symmetrical arrangements illustrated in Fig. 13.01 is that many were first used early in the history of cartography. Dahlberg (1962) and Keuning (1955) are the principal authorities on these early developments. Many of the star-shaped versions belonging to the class showing transpolar continuity date from the second half of the nineteenth century; those showing mid-latitude continuity became fashionable in atlas cartography even later. For example Bartholomew's *Kite*, *Regional* and *Lotus* projections all date from the 1930s and 1940s.

Continuity along a great circle: The use of gores

One of the earliest methods of producing a recentred map is still of major importance, for the equatorial continuity of repeated gores has been the method of producing a map to cover the surface of a globe ever since the early sixteenth century, and is the basis of the *Universal Transverse Mercator* (UTM) projection, which is not the most important integrated system used for topographical mapping in many parts of the world. The same arrangement is also used for the Russian version of the Transverse Mercator (*Soviet Unified Reference System or SURS*) which is used for mapping in many of the countries which do not, for political reasons, use the UTM.

A pattern of narrow gores which are contiguous along the equator is the printed source map for covering a globe. A typical example of this is illustrated in Fig. 13.02. The utility of this application has been described and illustrated in many twentieth-century textbooks and atlases. Moreover, the method is often used to demonstrate a way of making a plane map fit the curved spherical surface, for if the gores are made narrow enough the strips will fit the curved surface quite well. Globe-makers

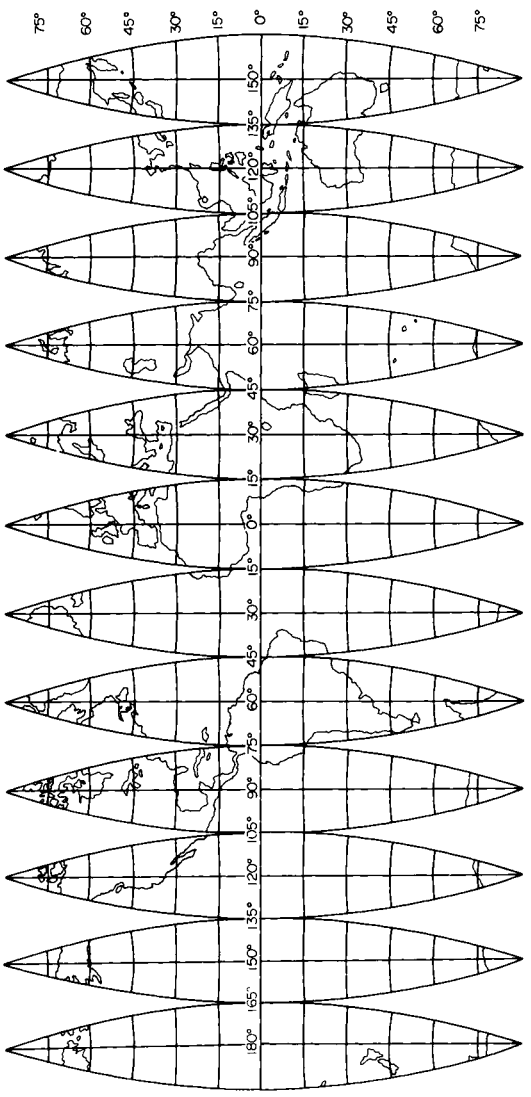


FIG. 13.02 An example of the conventional subdivision of the world into twelve gores, each corresponding to 30° longitude.

have produced the necessary cover in the form of gores which are less than 30° longitude or thereabouts in width, so that 12 gores or more are needed to complete the world globe. For this purpose each gore comprises curved meridians which are symmetrical about the rectilinear central meridian, and the parallels are also curved. However, Dahlberg has also identified some early examples which were clearly unsuitable for covering globes because the gores are of unequal width, are simply too wide, they are not tall enough to reach the geographical poles of the globe or the parallels are represented by straight lines. All of these factors would make it impossible to fit the map to a curved surface without producing an irregular appearance. Therefore these were originally intended as conventional maps to appear in atlases.

The modern use of gores is conceptual rather than actual, for it is sufficient to understand the organisation of each zone of the UTM or SURS, both described in Chapter 16, without having to make a map of it. In both systems the world between latitudes 80°(or 84°)N and 80°S is mapped in a series of gores which are only 6° in longitude wide. Each gore has it origin at the point where the central meridian intersects the equator. The central meridian is regarded as a straight line, but all other meridians are curves which are concave towards the origin. In the UTM a further modification has the effect of locating two lines of zero distortion which lie some distance either side of the central meridian. In the Soviet Unified Reference System no such modification is applied, so that the central meridian of each gore is the line of zero distortion.

Polysuperficial projections

Some writers, for example Goussinsky (1951) and Richardus and Adler (1972), employ a threefold subdivision of the *coincidence* of the plane of projection into tangent, secant and *polysuperficial.* The first two were described in Chapter 5. The last of these categories corresponds to the use of many planes, cones or cylinders in such a way that they create a number of separate projections having more or less identical characteristics. The archetypal polysuperficial projection is the polyconic projection, the geometry of which may be imagined as a nested succession of cones whose vertices, in the normal aspect, lie on the continuation of the polar axis. The appearance of a projection of only part of spherical surface may be imagined from examination of the central part of the world development of this projection in Fig. 5.02. The salient features of a polyconic projection of part of the world are:

- the central meridian is rectilinear,
- all other meridians are concave towards the central meridian,
- each parallel is circular but has its unique centre so that the parallels on the map are not concentric arcs.

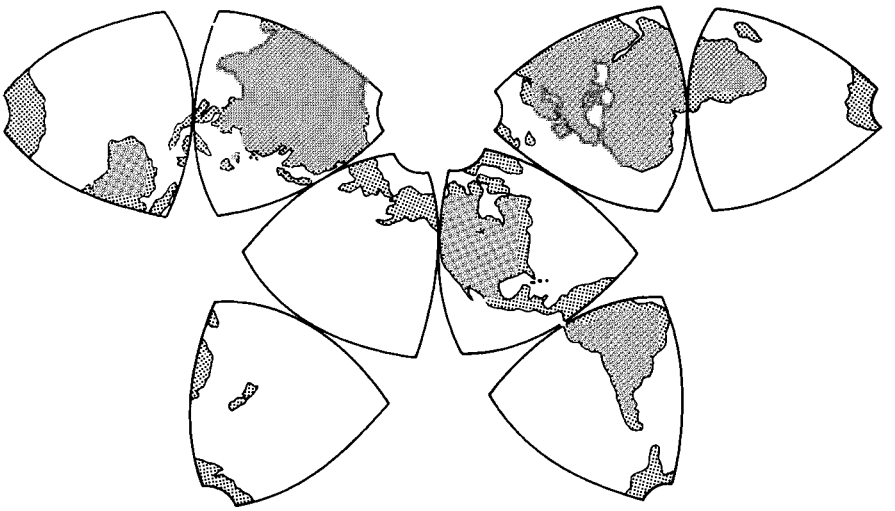


FIG. 13.03 An eight-lobed representation of the world formed from the polyconic projection illustrating the rolling fit between the lobes. This version has been called *Murphey's Butterfly projection* by Bunge (1962), but it differs little from one originally described by Cahill (1909).

Figure 13.03 illustrates the outlines of a polyconic projection for a large portion of the earth, extending from the equator to latitude 80° and having a width of 90° of longitude; 45° on each side of the central meridian. Eight such projections provide a map of the whole world as far as 80°N and S. These will match one another along the common bounding meridians and at one point on the equator. However, the reverse curvature of the equator between two contiguous lobes in opposite hemispheres or between corresponding points on adjacent bounding meridians limits matching to only one pair of tangential points at a time. However, different point pairs may be chosen, simply by altering the orientation of one lobe to the other. This is equivalent to rotating one lobe about the other, and is therefore known as a *rolling fit*. One way of fitting the eight lobes together is by setting the four northern hemisphere lobes so that they touch in latitude 40°, each southern hemisphere lobe meeting its northern counterpart on their common central meridians. The result is clearly a symmetrical recentred map of the world. We have described this map in some detail because its demonstrates that in the broad view this group of polysuperficial projections is, in effect, a recentred or interrupted map. More commonly, of course, we encounter it in the form of sheet maps. A version of the polyconic projection used to be the standard base for the International Map of the World at 1/1 000 000 scale (IMW) which covers the land areas of the world as a uniform sheet series. Before 1962, when a special UN conference decided that the map may also be compiled

on the Lambert conformal conical projection, each map was compiled to its unique Polyconic projection. Because of the property of this projection, which we have described and illustrated in a more exaggerated form, the phenomena of reverse curvature of the bounding meridians and the property of the rolling fit indicated the separateness of each map. Although the amount of curvature is small, and can often be ignored in mounting pairs of IMW sheets on a board or wall, the cumulative effect of the rolling fit makes it virtually impossible to fit more than four adjacent IMW sheets to one another without obvious overlap or gaps.

Another example of polysuperficial mapping at the larger map scales has been the use of the *polyhedric* or *trapezoidal projection* with sundry less familiar names, such as *Müffling's projection.* This is a form of pseudocylindrical projection with rectilinear meridians and parallels for the sheet boundaries, and it has been used by a number of countries, particularly in central and eastern Europe, as the base for their early map series. In this form the country is effectively subdivided into a series of trapezia; these being the four-sided rectilinear figures formed by two meridians and two parallels. The lengths of these sides are generally made to correspond to the arc distances along each graticule element.

Polygnomonic projections are a subset of the polysuperficial projections which employs the technique of repetition to overcome the difficulty inherent in the Gnomonic projection; namely that it can only be used to map smaller portions of the world. Thus the Gnomonic projection (No. 7 in Appendix I) which has the special property that all great circle arcs are shown by straight lines. However, use of the Gnomonic projection is restricted by the limitation that it can only be used to depict less than one hemisphere. For practical purposes the use of it is more or less limited to a maximum value of $z = 60^\circ$ from the origin. It is, however, possible to repeat the representation of a single unit of the Gnomonic projection which may be artificially bounded by the sides of a triangle, square, pentagon or other geometrical figure, and to fit these units together, thereby creating a polygnomonic projection which extends over a much larger part of the earth's surface. The best-known version, shown in Fig. 13.01, is the map comprising six components, these corresponding to the faces of a cube whose planes are tangential to the earth at the centre of each component. Such a map proved quite useful in the early days of long-distance flying when it was necessary to follow great circle routes across the Atlantic or Pacific Oceans and to fly from western Europe or the USA to Australia, South Africa and South America. Although these early uses are no longer of any practical importance, these patterns of 'cartographic wallpaper' are not necessarily to be relegated to pictorial design and the creation of suitable logos for commercial airlines. Tobler and Chen (1986) have hinted at the possibility of using such designs as means of storing GIS data. There is research to be done in this field.

Combined with some ingenuity in the location, shape and extent of each facet, it is possible to produce polysuperficial projections printed on card which may be cut out, folded and pasted to create a model of the earth corresponding to a geometrical solid. Because such devices depend for their success upon the publicity surrounding their invention and marketing, the public become aware of some and never hear of other examples. Thus the so-called Dymaxion Globe by Buckminster Fuller is well known in this field largely owing to the clever publicity which has always surrounded it. Such toys are also well protected by patents.

This application of polysuperficial projections may be used in other ways. For example, in the definitive study of conformal projections based upon elliptical functions, Lee (1976) has produced a variety of conformal maps of the sphere or a hemisphere enclosed by different plane and solid geometrical figures. Those which are based upon the faces of geometrical solids have the repetitive character of the polygnomonic projections although, of course, their special properties are quite different. Lee's version of a conformal map of the world based upon the icosahedron, and which therefore is composed of 20 separate triangular facets, has been used most effectively by Eckhardt (1983) to illustrate sea-surface topography derived from Seasat altimetry data. One of the most interesting of these projections is Lee's *Conformal Tetrahedric projection*, also described in detail in Lee (1965, 1973, 1976), in which the surface of the

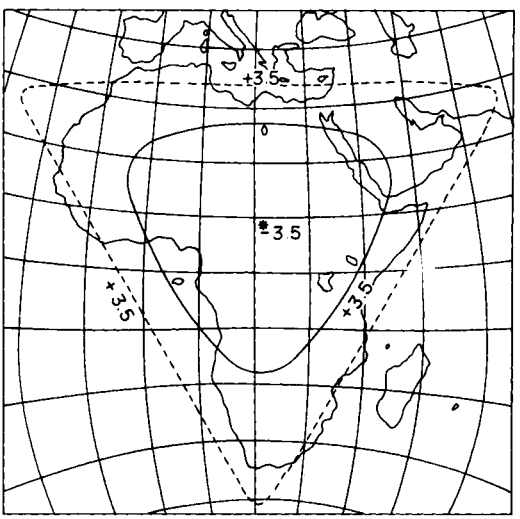


FIG. 13.04 Conformal map of Africa based upon a portion of Lee's Conformal Tetrahedric projection and described by him in Lee (1965, 1973, 1976). Isograms are for scale errors of −3·5%, 0 and +3·5%.

sphere is mapped to four equilateral triangles. Apart its use for world maps, this remarkable projection has great potential use for depicting certain continents and combinations of continents. For example, Fig. 13.04 illustrates Lee's example of a map of the continent of Africa based upon the conformal tetrahedric projection. The distortion isograms have a characteristically triangular pattern and show that the scale errors for virtually the whole continent lie within the range +3·5% through −3·5%, which are the same limiting conditions examined in Chapter 12 for the conformal map of Hispanic America.

Asymmetrical recentred maps

To most map users the interrupted map projection is a world map in which a series of lobes are connected to one another along the equator. Most of them have been derived from one of the pseudocylindrical projections such as the Sinusoidal or Mollweide's projection. Each of these lobes is mapped continuously according to the principles of the parent projection. Generally the lobes are of different width and are offset to provide optimum cover of the world land masses so that such maps are asymmetrical.

The object of such interruption, as we have seen, is to reduce the deformation towards the edges of the world map by limiting the size of each lobe to only a small longitudinal extent either side of its central meridian. Any of the pseudocylindrical projections and a number of polyconic projections may be treated in this fashion, but most of the examples which have actually been used have been based upon equal-area pseudocylindrical projections. It follows, therefore, that it is the maximum angular deformation, ω, which is most often used as a measure of deformation on these maps.

Figures 7.04, p. 132 and 13.05, p. 276 have already illustrated the technique as this has been applied to the Sinusoidal and Eckert VI projections. However, there is a multitude of other projections which may be treated similarly. Generally the world map is subdivided into five lobes; two in the northern hemisphere so that South America, Africa and Australasia are mapped separately. This usually means the choice of central meridians in longitudes 20°E (or 80°), and 90°W for the northern hemisphere; in longitudes 30°E, 130°E and 70°W for southern hemisphere. For mapping the oceans the corresponding central meridians are chosen in longitudes 60°E, 30°W and 170°W (northern hemisphere) and 90°E, 20°W and 140°W (southern hemisphere) to create a six-lobed map of the oceans.

In each case the choice of the positions, number and width of the lobes is a matter of compromise between the marginal deformation if they are made too broad, and the discontinuities caused by the gaps if they are too

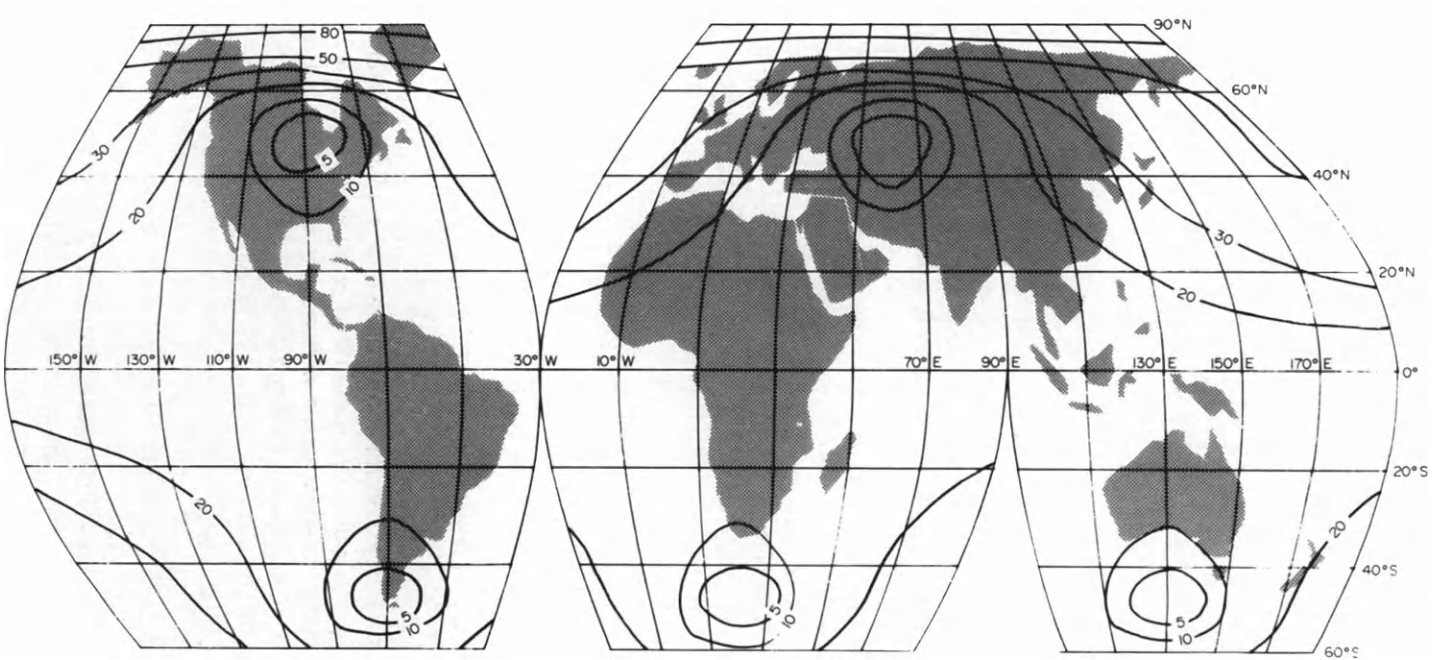


FIG. 13.05 A recentred version of the Eckert VI projection (No. 25 in Appendix I). This is an equal-area pseudocylindrical projection in which the meridians are sinusoidal and the geographical poles are represented by straight lines which are one-half of the length of the equator. In addition the world map illustrated is recentred, having central meridians in longitudes 90°W, 70°W, 20°E, 60°E and 130°E. The isograms are for maximum angular deformation (ω) at 5°, 10°, 20°, 30°, 50° and 80°.

TABLE 13.01 *Robinson's analysis of the mean values of maximum angular deformation (ω) within the land areas of certain pseudocylindrical equal-area projections have been interrupted in the same way*

Projection	Mean deformation of land area		Approximate percentage decrease
	Uninterrupted	Interrupted	
Sinusoidal	35°36′	19°20′	46
Flat polar Quartic			
Authalic	31°00′	18°19′	41
Truncated Sinusoidal	29°12′	18°30′	37
Mollweide	29°00′	17°00′	41
Truncated elliptical	27°57′	18°00′	36
Truncated parabolic	26°59′	18°00′	38
Eckert IV	24°45′	17°54′	28

numerous. At one extreme we have the discontinuous map created by a string of contiguous narrow gores; at the other there is increasing angular deformation in a large lobe such as that illustrating Eurasia in Fig. 13.05.

Early descriptions of such maps provided somewhat subjective statements about the appearance of different parts of each map with no attempt at quantitative evaluation of the merits of the parent projection or an interrupted version of it. Robinson (1953) undertook such a study using a form of analysis which he had already used to good effect with other world maps for use in atlases. The procedure involves:

1. outlining the land area of major interest (Antarctica is excluded);
2. plotting the values and drawing isograms of equal maximum angular deformation;
3. measuring by planimeter the land area enclosed by succeeding isograms;
4. deriving the mean of the angular deformation for the land from a graph of deformation plotted against area.

In order to make an adequate comparison of the different interrupted projections, the pattern of lobes should be the same for each. Robinson made a comparative study of seven different pseudocylindrical projections. The results of this work are given in Table 13.01.

Recentred composite maps

The example of an asymmetrical composite projection examined here offers an interesting variation from the conventionally recentred pseudocylindrical projections. It also demonstrates how particular ideas may resurface periodically at intervals of a generation or so.

Towards the end of the nineteenth century, Frye published an un-

familiar version of the distribution of the continents which originally appeared as a rough sketch in his book *The Child and Nature*, published in 1889. This was more fully developed in his *Complete Geography*, first published in 1895 and subsequently reissued many times and in many countries. Although the construction is not described by Frye, it appears to consist of the land hemisphere on the oblique Azimuthal equidistant projection with two extended lobes of different construction. The map next appeared as a rough sketch in Mackinder's *Britain and the British Seas* (1902) without any acknowledgement of the source. Interest in it was revived by Fawcett (1949), more than 40 years later. Fawcett was clearly unaware of Frye's work but states that it was the illustration in Mackinder's book which stimulated him to attempt production of a suitable graticule. He used the Azimuthal equal-area projection as the basis of his map. He described and illustrated two versions which here appear as Fig. 13.06(a) and (b). The first is based on the normal aspect of the Azimuthal equal-area projection; the second is based upon an oblique aspect of the same projection centred on London. It is the second which corresponds closely to the original conception of Frye.

In Fig. 13.06(a) the South Pole is the origin and the sole point of zero distortion. The whole of the southern hemisphere therefore corresponds to a normal aspect Azimuthal equal-area projection and is therefore identical with Fig. 10.02, p. 201 as far as the equator. The northern hemisphere is depicted by means of three lobes. If the map is intended to show maritime distributions, the lobes are chosen to accommodate the Pacific Ocean, the Indian Ocean and the Atlantic Ocean. The meridians used are: 160°W for the Pacific, 60°E for the Indian Ocean, and 40°W for the Atlantic. Although these might be achieved by extending the Azimuthal equal-area projection to the North Pole, the deformation in high northern latitudes becomes excessive. Consequently Fawcett devised the following empirical method of construction. Each lobe has a rectilinear central meridian along which the separation of the parallels is the same as occurs in the normal aspect Azimuthal equal-area projection. However, the curvature of these parallels is reversed so that they appear to continue the southern hemisphere mapping. In order to maintain a semblance of equivalence the meridians have to be arranged in such a way that the area contained within a quadrangle in the northern part of the map is equal to the area of the corresponding quadrangle in the southern hemisphere. This leads to the difficulty that the northern hemisphere meridians are not longer rectilinear, but are curved.

The oblique aspect map is based upon the Azimuthal equal-area projection with its origin near London ($\varphi_0 = 50°\text{N}$, $\lambda_0 = 0°$). It has two projecting lobes, one having an axis passing through South America; the other through Australasia. Fawcett claimed to have constructed this by means of graphical transformation from the source, Admiralty Chart

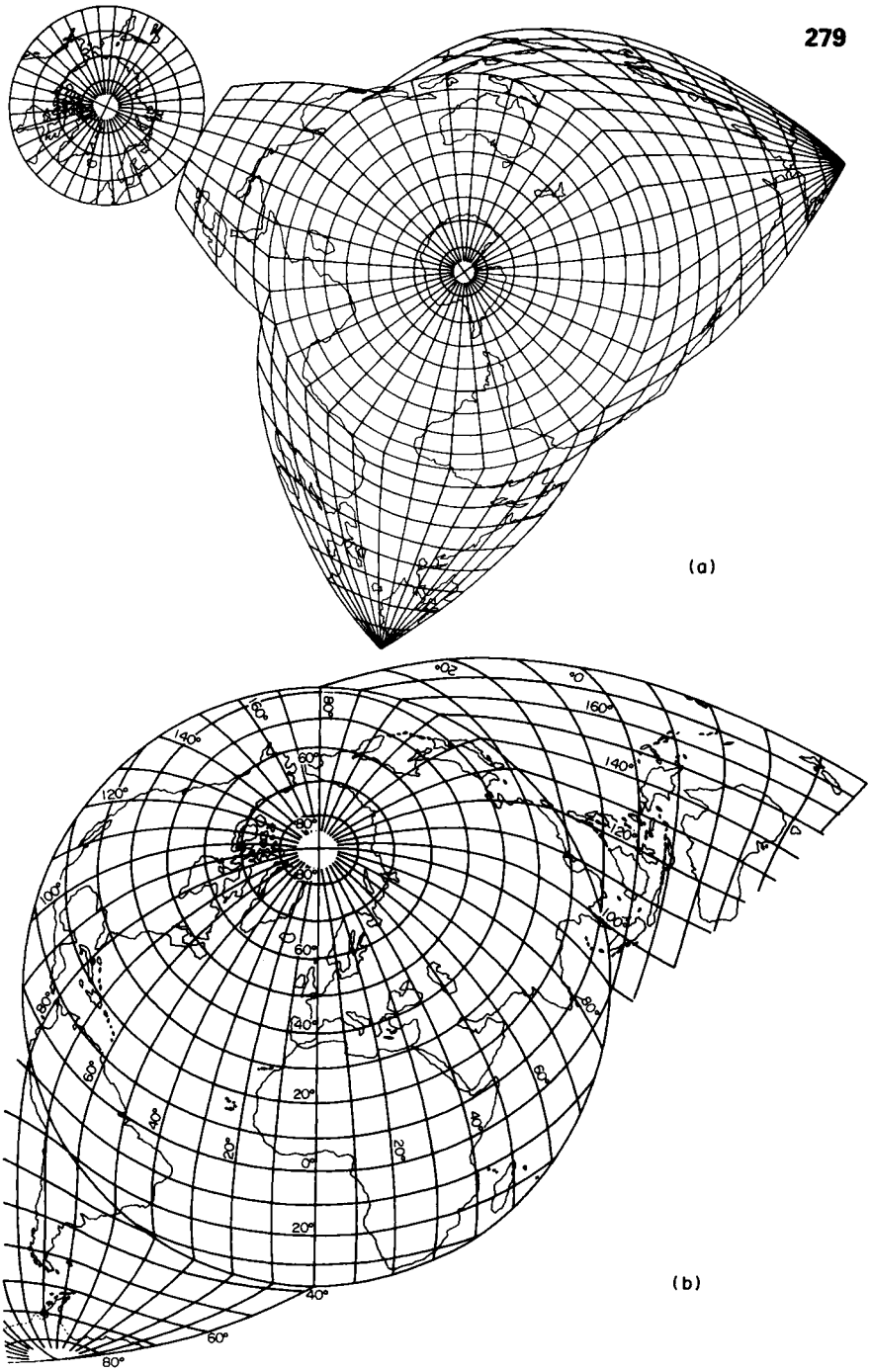


FIG. 13.06 (a) The normal aspect of Fawcett's composite equal-area projection as modified by Fisher and Lockley (1954) for the *New Naturalist* volume on *Seabirds*. The central meridians of the lobes are selected to give oceanic continuity. Fisher's improvement on the original is to replace the circular representation of the Arctic (or Antarctic in a map centred on the North Pole), like a small wheel in each lobe, with a point pole. (b) The oblique aspect of Fawcett's composite

5085, which is based upon an oblique aspect Azimuthal equidistant projection centred on London, but his description of the construction of the two southern hemisphere lobes is far from clear.

The only use even made of Fawcett's work appears to have been a series of maps which were published in volume No. 28, *Sea-birds*, by James Fisher and R. M. Lockley (Fisher and Lockley, 1954), this being one of the Collins *New Naturalist* Series. They demonstrated the particular value of the projection in a different form to illustrate southern hemisphere distributions as in Fig. 13.06(a) and in the oblique version to show palaearctic distributions of species. Fisher's principal contribution was to modify the representation of the more distant geographical pole by a series of points, whereas Fawcett had left them as three separate representations of Antarctica (or the Arctic Basin) like little wheels at the end of each gore.

At this stage the potential of the projection was again forgotten for another 40 years until another version of it was announced by M. C. Jackson (1988, 1990). He, in turn, was quite unaware of the earlier work, but hit upon the idea of using the Azimuthal equidistant projection for a map which is very similar to Fawcett's second projection and is, in other words, close to the map which Frye originally intended. Jackson's model comprises three lobes to accommodate the southern hemisphere land areas, but by better choice of the origin for the northern hemisphere map it may be converted into the simpler two-lobe version. Then the only distinction is that Fawcett's map is equal-area throughout whereas Jackson's is the typical 'good compromise' map afforded by the use of the equidistant projection. Because of the different special property, this justifies separate classification of Jackson's map even if it used similar lobes to Fawcett's map.

Condensed projections

We must also distinguish a third technique of deliberately breaking the continuity of the mapped surface which is simply to remove unwanted parts of a map, such as the oceans from a world map intended to show terrestrial distributions. For example, a map intended to show terrestrial distributions in both Old and New Worlds may be designed so that the majority of the Atlantic Ocean is not shown. The usual way of doing this is exemplified by some of the maps published in the *Oxford Atlas*, where a pair of maps were compiled for different origins on the equator and were printed in juxtaposition with a geometrical boundary through the Atlantic Ocean. A similar arrangement was used by Barney (1980). The result may be called a *Condensed projection*. The gaps do not, of course, have any effect upon the distortion patterns of the maps, which remain wholly independent of one another. The only benefit bestowed by the

technique is to allow the map maker to show the land masses at a larger scale than would otherwise be possible on a particular size of paper. The condensed map should, however, possess some kind of continuity with the original source. Each element should be at the same scale and there should be some continuity in presentation such as placing the equator, or a meridional axis on the same straight line. In this respect the condensed map differs from a conventional map having *insets*, which are usually outlying parts of a country (Alaska and the Hawaiian Islands on maps of the USA; the Shetland and Orkney Islands on a map of Britain) and which need to be shown in a conveniently blank area occupied by the sea at a larger scale (Hawaii) or smaller scale (Alaska) than the main map. Strictly speaking the word 'interrupted' is more appropriate than 'condensed' for such maps, so that the first might well be applied to this convention only. However, this word is almost invariably used to describe recentred projections so that it would be confusing to attempt to change the terminology. Although this may help to depict a distribution by showing it at a larger scale than would be possible on a continuous world map, it does absolutely nothing for the distortions on the parts of the map which have been retained. Moreover, the interpretation of such maps may be misleading if the unobservant user fails to appreciate the size of the gaps between different parts of the map. Misunderstanding of this convention leads to stories like the one about the travel agent who insisted with a faith no argument was capable of shaking that the Shetland Islands were only 50 miles from Aberdeen, for he had measured this himself on a map. He had failed to realise that an inset map of the Shetland Islands, located for convenience in the Moray Firth, was not in its correct spatial relationship to the map of the mainland.

Map projections which introduce deliberate distortion

Thus far we have concentrated wholly upon the reduction of the inherent distortions of map projections. The goal has been to reduce the map to one of constant scale through the removal of the distortions over as much of the map as is possible. It is, of course, an unattainable goal.

We must now enquire whether there is any advantage in making some parts of a map more exaggerated than others and, if so, by what means this may be achieved. Such exaggerated diagrams are often called *cartograms*, and have been dismissed as being mathematically inferior to true maps. In the early days of their use this was to a certain extent true, although it would be reasonable to suggest that Frye and Mackinder's version of the world map described earlier would have been described in this fashion. It is really a matter of adequate description which converts the caricature into a useful map, even if this exaggerates some parts rather than others.

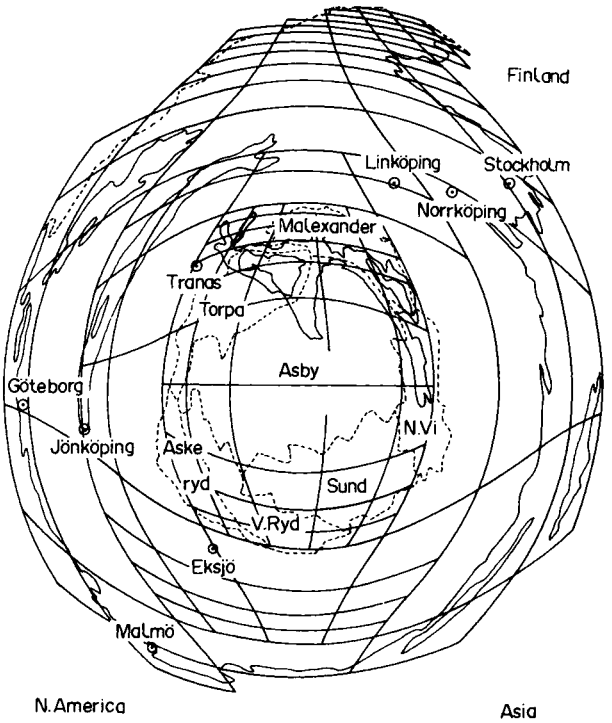


FIG. 13.07 Hagerstrand's famous diagram based upon Kant's Logarithmic Azimuthal projection.

The Logarithmic Azimuthal projection

One of the first attempts at quantitative representation of a non-metric space was a famous map by Hagerstrand for whom Edgar Kant devised the *Logarithmic Azimuthal projection* described by Tobler (1963). This was intended to express the extent of emigration from a particular locality in Sweden which needed to indicate quite minor movements into neighbouring towns or districts, as well as emigration to distant places. The rate at which the radial scale falls off from the scale at the centre is still more flexible when using a logarithmic or other specially designed azimuthal projection. One form of a Logarithmic Azimuthal has a radius

$$\rho = R_b \ln(1+\sigma z)/\ln(1+\sigma b) \qquad (13.01)$$

where ρ is the radial distance from the centre to a point at distance z from the centre, R_b is the radius of the outer circle bounding the map, b is the map range up to 180° in the same units as z, and σ is an arbitrary constant chosen to obtain the desired enlargement near the centre. In short the scale of the map is much greater at the centre of the map than it is near

the edges. As σ approaches zero, the plot approaches an Azimuthal equidistant projection.

The Hyperboloid projection of the Falk Town maps

The next example of how maps which distort excessively may have other practical uses may be recognised in the production of two maps for tourists. Many towns have crowded central areas, corresponding to 'the old city', 'the mediaeval town' or even 'downtown'. Further afield the density of housing becomes less, and in the outer suburbs there may be large open spaces between clusters of dwellings. This means that a typical large- or medium-scale map covering the whole town is either too large a scale for the suburbs, requiring a very large piece of paper or several map sheets to cover these parts of it, or the map is of too small a scale to show the congested inner city adequately. The traditional method of overcoming this is to produce a comparatively small-scale map covering the whole of the town including the outer suburbs, with inset large-scale maps of the congested inner-city areas. Almost invariably the area of particular interest to the user lies just beyond the large-scale inset so that it is necessary to check backwards and forwards from one to the other. A most ingenious solution to this problem was proposed in the 1950s by the firm Falk Verlag in Hamburg, who started to produce variable-scale town and city plans at this time. These are described as being based upon a *Hyperboloid projection* with a kilometre grid. Typical scale variations on the maps are 1/15 000 down to 1/30 000, or the scale at the centre of the map is double that at its edges or corners. The change in scale is continuous so that there are no sudden breaks in the map. The kind of deformation which is present is obvious from the curved grid lines, which are everywhere convex outwards. Since these maps are produced commercially and, until the early 1970s had no competitors, the publishers did not publish any information about the projection or its derivation. The following solution was provided by Doytsher for Kadmon (1975).

The Hyperbolic projection

Assuming a linear change in scale S radially in any direction from the centre of the map results in a projection in which a square grid is represented by parabolae. If we denote the centre of the map as having coordinates (X_0, Y_0) where the scale is S_1, and at one of the other extremities of the map (e.g. the SW corner) of the map where the coordinates are (X, Y) and the scale is S_d, we have, assuming linearity

$$S_d = S_1 + (S_2 . S_1/m) . d \qquad (13.02)$$

where d is the distance from (X_0, Y_0) to (X, Y) and m is the distance

between the centre of the map and the point at which the scale is S_2. Obviously

$$d = [X^2+Y^2]^{1/2}$$

on a plane map. Therefore, putting

$$c = (S_2 . S_1)/m$$

and

$$S_d = S_1 + c . [X^2+Y^2] \tag{13.03}$$

Also the particular scale is

$$S_d = dx/dX$$

$$dX = S_d . dx \tag{13.04}$$

we have

$$dX = S_1 . dx + (c . [X^2+Y^2]^{1/2}) . dx \tag{13.05}$$

Therefore

$$X = \int S_d . dx$$

$$= \int S_1 dx + c \int [X^2+Y^2] . dx \tag{13.06}$$

$$= S_1 x + c . [X^2+Y^2/X^2]^{1/2} . x^2 \tag{13.07}$$

Putting

$$c_x = c . [X^2+Y^2/X^2] \tag{13.08}$$

we re-write

$$X = S_1 . x + c_x x . x^2 \tag{13.09}$$

and thus obtain a quadratic equation in x

$$c_x . X^2 + S_1 . x . X = 0 \tag{13.10}$$

The solution of this is

$$x = -S_1 \pm [S_1^2 + 4c_x . X]/2c_x \tag{13.11}$$

Similarly, if we denote

$$c_y = c . [X^2+Y^2/Y^2]^{1/2} \tag{13.12}$$

we obtain

$$y = (-S_1 \pm [S_1^2 + 4c_y . Y]^{1/2})/2c_y \tag{13.13}$$

Only the positive roots are required. Thus we find that the square grid is transformed into parabolae, while scale 1/S, the inverse of linear reduction, is hyperbolic.

The next problem is to transform map detail from an ordinary map of the town of uniform scale to the extremely variable scale of the Hyperboloid projection. This posed formidable difficulties in the days before digital mapping, because of the large and continuous changes in scale which had to be incorporated into the new map. However, this kind of work is ideally carried out using digital methods and this was the method used by Kadmon for his town map of Jerusalem illustrated in Fig. 13.08.

Such maps retain all topological properties as well as true angles at the

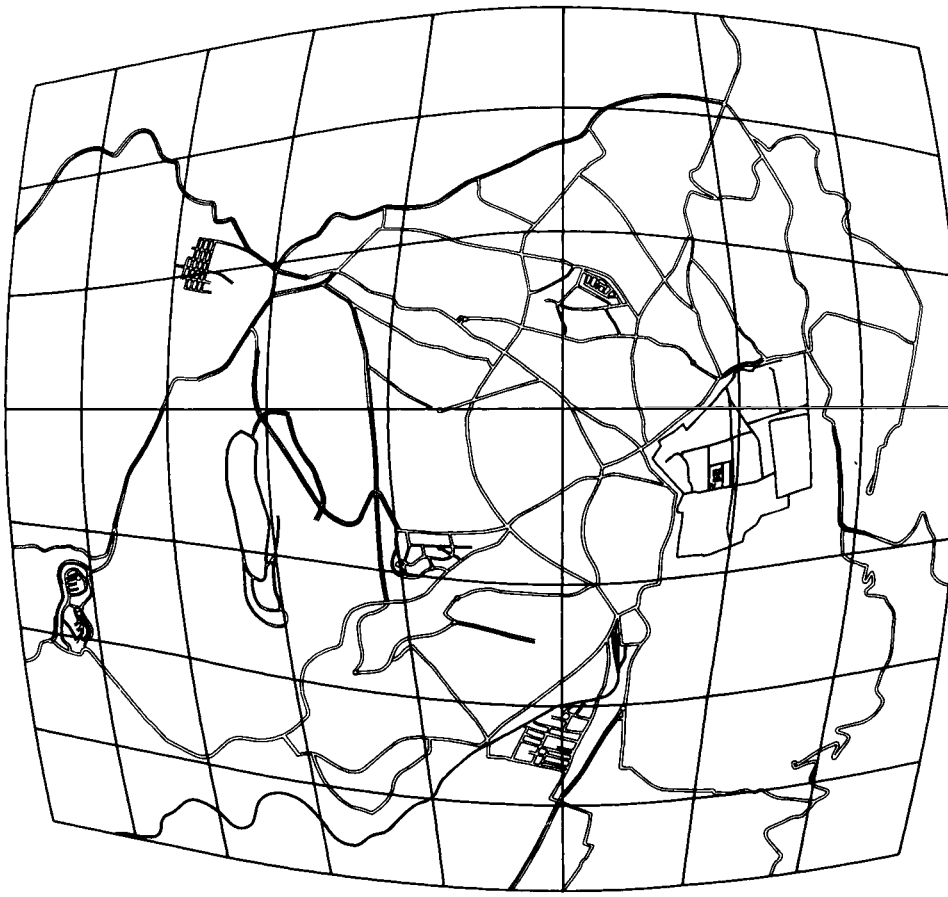


FIG. 13.08 A variable scale town map of Jerusalem calculated by the Doytsher–Kadmon method and plotted on a Gerber flat-bed plotter. Note the shapes and sizes of the grid which are uniform squares on a conventional map. (Source: Kadmon, 1978.)

centre. Kadmon has called them *Azimuthal cartograms*, and has shown that there are many interesting applications other than provision of convenient tourist maps for historic towns. For example, an interesting application is in the field of road transport. If the average traffic speed in a city centre and its increase with radial distance (with decreasing congestion) is known, map scale can be made to represent driving time, e.g. in min/km, and average driving times in minutes can be read directly from the map. Moreover, the computer/plotter can easily produce separate maps for different times such at mid-day, under rush hour conditions, etc.

The concept used to design the Hyperbolic projection may be generalised to accept the different variable values. A detailed derivation of the maping equations needed to transform the (x, y) coordinates of a conventional map into the (x′, y′) system appears in Kadmon and Shlomi (1978). This allows exaggerated, larger-scale, remapping of the source map around a single focus with scale tending to that of the source as radial distance increases.

Similar Russian work in this field has been briefly described by Vakhrameeva and Bugaevsky (1985). They illustrate an example of a world map of chemical production, in which the problem of depicting the great concentration of factory sites in Japan has been alleviated by locally increasing map scale.

The magnifying-glass effect

Snyder (1987d) has devised several new projections which have, in his words, a magnifying-glass effect. These are maps which appear to be viewed through a magnifying glass, having a much exaggerated scale within the circle or rectangle simulating the magnifier and a smaller scale beyond. The examples described by Snyder are based upon azimuthal projections, such as the Azimuthal equidistant and Azimuthal equal-area projections. Two techniques are presented. In the first the special properties of the projection are preserved for an inner and an outer portion with an abrupt change in scale at their boundary. In the second technique inner portion is a standard Azimuthal projection, but the radial scale beyond its boundary is gradually reduced until it is zero at the chosen outer limit. The simplest versions relate to circular boundaries corresponding to circular magnifying glasses, but the principles may also be applied to square or rectangular edges to the larger-scale portion of the map.

The first of Snyder's examples is illustrated in Fig. 13.09. This is Azimuthal equidistant version of the *magnifying-glass Azimuthal projection*. It is a true Azimuthal projection so that angles measured from the centre remain correct through the entire map. For the inner part of the map any distance measured radially from the centre corresponds to the

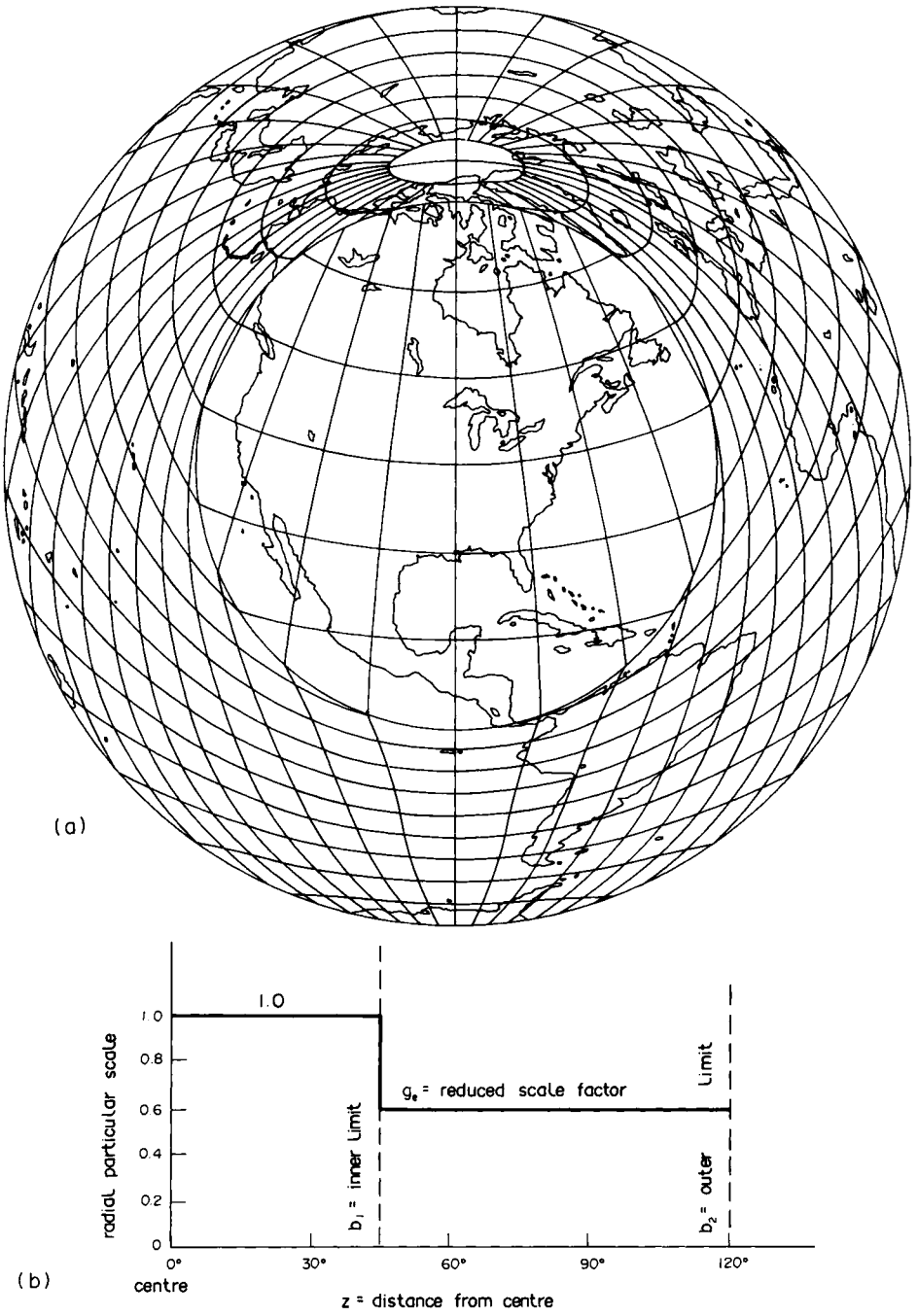


FIG. 13.09 (a) Snyder's 'magnifying-glass effect' applied to an Azimuthal equidistant projection; (b) graph of the particular scales for this projection.

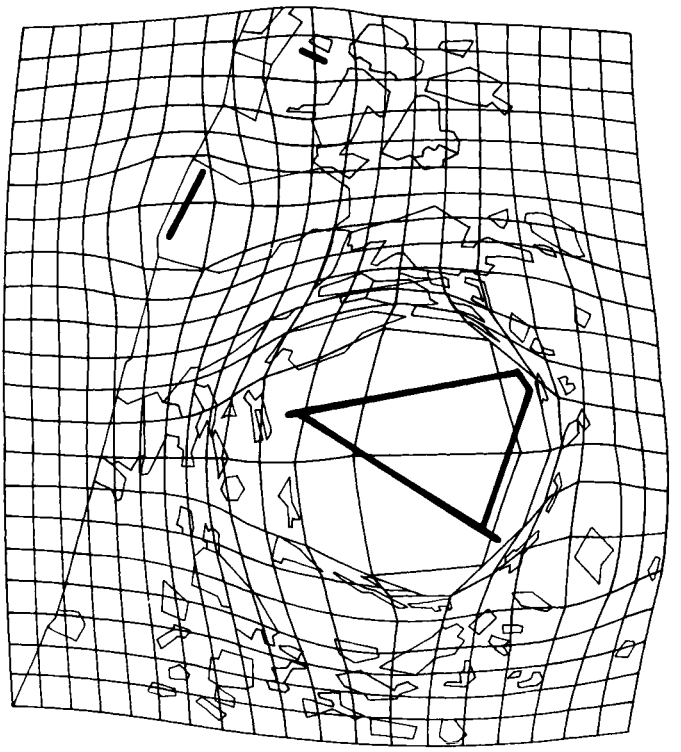


FIG. 13.10 Map showing the use of the Polyfocal projection to illustrate the distribution of noise in the vicinity of certain airports in Israel, in which the size of the airport is proportional to the measured output of noise.

great circle distance on the sphere at the stated scale of the map. Beyond the circle bounding this portion, the radial scale is reduced to a constant fraction of the stated map scale. Here it is 0·6 of the principal scale of the central part of the map. Figure 13.09(b) illustrates this distribution of the radial scales within such a map, and shows that the boundary is a simple step where the radial particular scale changes abruptly.

Snyder's second series of examples, which he called *tapered Azimuthal projections*, have a variable radial scale beyond the magnified area.

The Polyfocal projection

The next logical step from projections having only a single focus or point of origin which can be shown at a larger scale, is to a projection which has a number of arbitrary separate foci. Each may be allocated its own scale, proportional to some thematic variable (such as time and average traffic speed in the example already mentioned) and its individual 'friction decay' or 'friction function' around it. The scale at any other point which

is not one of the foci is made equal to the sum of the influences of all adjacent foci on this point.

Kadmon and Shlomi (1978) produced the Polyfocal projection to be used with the program POLYMAP to compute and plot the results. Returning to Kadmon's earlier example, of using the Hyperbolic projection to illustrate the influence of traffic congestion and journey time from a single centre, the combined influence of any number of traffic black spots may be studied in unison using such a technique.

An interesting practical example of the application of the Polyfocal projection which has been used by Kadmon and Shlomi (1978) and Kadmon (1983) is that of the noise level surrounding various airfields in the vicinity of Tel Aviv. Here the quantitative variable used to control map scale is that of noise in decibels.

CHAPTER 14

Projections for navigation charts

> Whiche waye too bee knowne is thys; Fyrste too consider by what poynte that the shippe hath made hir waye by and how fast and swiftly that the shippe hath gone, and to consider how often that the shippe hath altered hir course, and how much of that shee hath gone at euerye tyme, and then to consider all thys in your Platte or Carde, and so you may giue an neere gesse by what poynt or wynde it beareth from you, and also howe farr it is thyther.
>
> William Bourne, *A Regimen for the Sea* (1574?)

Introduction

We have already emphasised that navigation is one of the most exacting of all kinds of map and chart use. In order to explain why navigation charts must be based upon projections which satisfy certain combinations of special property, we must know something about the way in which charts are used. It is therefore desirable to interrupt the study of map projections to describe some of the principles and methods of navigation. The description which follows must, of necessity, be brief. Since our preoccupation is with the use of charts, we must ignore navigation techniques in which the chart use has secondary importance.

Navigation and pilotage

Navigation is the art of taking a craft from one place to another out of sight of land. *Pilotage* is the art of taking a craft from one place to another when land or navigation marks are in sight. The object of both is to ensure that the craft makes a safe passage from the place of departure to its destination, preferably along some predetermined *track* and within the time schedule allowed by timetable, available fuel supplies and similar constraints. These objects should be achieved without risking stranding or collision with rocks, sandbanks, wrecks or other shipping at sea; with high ground and other aircraft in the air.

Purposes of charts

An indispensable part of the equipment needed by the navigator is a chart, or more commonly a sequence of charts, covering the route to be followed. Charts fulfil three requirements:

- *They provide information about the nature and position of hazards to navigation.* These include shallow water and submarine obstacles to be shown on the nautical chart; high ground and overhead obstacles to be shown on aeronautical charts.
- *They provide information about the availability and identification of aids to navigation.* These include marine lights and buoys on nautical charts and radio direction-finding aids, such as beacons and radio ranges, on aeronautical charts. Where such aids are available, both marine and air charts may show the network denoting lines of constant instrumental readings which are used to fix position in the various kinds of hyperbolic navigation systems (e.g. *Decca*, *Shoran*, *Loran*, *Consol*). These are known as *lattice charts.*
- *A chart is the base upon which the graphical work of navigation is done.* It is this function of the chart which is most closely associated with the choice of projection and which is therefore studied here.

For the present we may regard the procedures of both marine and air navigation as being identical. This was true of the early days of flying, when the techniques of air navigation evolved from those already in use at sea. This similarity still exists in the navigation of slow piston-engined aircraft which are not equipped with the modern sophisticated avionics systems, but the methods of navigating high-performance aircraft have changed as flying speeds have increased, simply because there is insufficient time for the navigator of a jet aircraft to solve problems graphically on a chart. These have to be done by analogue or digital computation rather than by pencil, ruler and dividers. We will comment briefly how these changes in technique have influenced aeronautical chart design after considering the fundamental similarities which exist between marine and air navigation and the ways in which a chart is used.

Stages in a flight or voyage

Any voyage or flight may be divided into three separate stages. The first and last of these are the periods immediately after departure and prior to arrival. At such times a ship is close to land and is probably confined to a navigable channel which is crowded with other shipping. During the intermediate *en-route* stage the vessel is out of sight of land, in deep water and the risk of collision is much less. In a navigable channel the facilities for fixing position are usually frequent and reliable. In the open

ocean, aids to location are generally much less satisfactory, and the skill of the navigator lies in making a correct interpretation of a variety of data to ensure that the craft maintains the proposed track according to the intended timetable.

In flying the same distinction can be made in flying between those periods just after take-off and shortly before landing, when the position and height of the aircraft is ordered by flying control, and the *en-route* period of flight when surveillance from the ground is less stringent. However, as the volume of air traffic has increased, and with it the need to maintain safe clearance between aircraft, the amount of navigation to be done by the crew has declined, and has been replaced by precise instructions from air traffic controllers. For example, the movements of aircraft operating over Europe and North America are now almost wholly controlled from the ground. Under these circumstances the navigator has no more opportunity to exercise his knowledge, skill and judgement during the *en-route* stage of flight than has the mariner sailing up the Manchester Ship Canal. Because we are primarily concerned with the study of the projections which are used for navigation charts, rather than the other aspects of chart content and design, we limit the present discussion to methods of chart use during the *en-route* stage, when the job of the navigator is to keep the craft on track, on time and to avoid 'getting lost'.

Dead-reckoning (DR) navigation

Certain information about the performance and movements of the craft can be measured on board.

Direction may be measured by compass, directional gyro or gyro-magnetic compass. By using these the helmsman or pilot can steer the craft on the required heading. This course can usually be maintained within $\frac{1}{2}^\circ$ of the intended direction. Bearings can be measured to similar order of accuracy. It is important to stress that this order of accuracy is lower than is needed in surveying and gunnery, and influences some of the assumptions which can be made about the properties of certain projections. *Speed* or *distance travelled* may also be measured instrumentally. In an aircraft, the *airspeed indicator* measures the airflow past the wings and fuselage. At sea, the distance travelled can be measured by a *patent log* towed astern, though this is also being replaced by meters recording speed, similar to the airspeed indicator. When the necessary corrections have been applied to take account of the influence of the environment, measurements of speed are likely to be accurate within 1–3%.

However, these measurements of direction and distance are relative because the movement of the water or the air also affects the movements

of the craft. The course of a ship or aircraft does not necessarily correspond to the desired track, and the ground speed is not the same as the airspeed. The distance recorded by patent log will underestimate the actual movement of the vessel if there is a following sea.

The absolute movements of a ship or aircraft over the earth's surface are exceedingly difficult to measure continuously or reliably. Before the *doppler navigator* was introduced in the middle 1950s it was impossible to measure directly the track and ground speed of an aircraft once it had passed out of sight of the ground. *Inertial navigation systems* now make possible continuous recording of position in any craft under any circumstances, for example, on board a submarine operating beneath the pack-ice of the Arctic Ocean. However, the size, cost and sophistication of the equiment required preclude their use for many purposes, and in the absence of such instruments the navigator still has to determine the unknown quantities of track and ground speed by indirect methods.

In order to show how these may be obtained by plotting on a chart we refer specifically to the navigation of a piston-engined aircraft with a cruising speed within the range 100–150 knots (115–173 mph or 185–278 kph). We use this example to represent the problems of DR navigation in their most acute form, because the influence of the wind is important. The airspeed of a jet aircraft is much greater so that the influence of the wind is proportionately less. The speed of a ship is slow, but the displacement by the sea is also small.

Triangle of velocities

The relationship between the measured and actual movements of an aircraft are characterised by the *triangle of velocities* (Fig. 14.01) in which the three sides of the triangle are vectors having both direction and length. Thus the angles of the triangle are represented by the differences in direction of the adjacent sides, and the lengths are proportional to speed. One side of this triangle is composed of known quantities. These are the course and airspeed of the aircraft, obtained from measurements made on board using the instruments already mentioned. Another side of the triangle is represented by the wind velocity, or the vector comprising wind speed and direction. This is the force displacing the aircraft during flight. The third side is the resultant of these components, representing the track and ground speed of the aircraft. The interior angle of the triangle formed between the course and track vectors is known as the *drift*. It may be imagined that as the length of the side representing airspeed is greatly increased, as for a jet aircraft, but the wind speed remains constant, the drift becomes less. Since one side of this triangle is always known, the triangle may be solved if another side can be determined. For example, if we can determine the wind velocity we may calculate track and ground

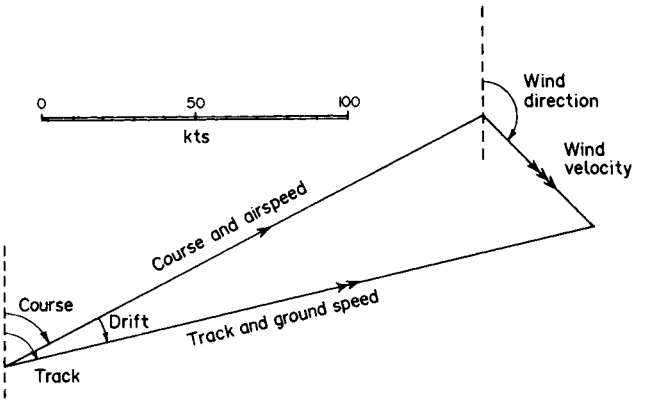


FIG. 14.01 The triangle of velocities in air navigation. The angles of the triangle are formed by observed directions and the lengths of the sides of the triangle are proportional to speed according to the scale provided. Hence this diagram illustrates the effect of a north-westerly wind of 50 knots upon an aircraft flying with a heading of 062° and airspeed 177 knots.

speed. Conversely, if we can *fix* the position of the aircraft, we can measure the track and calculate the ground speed from the chart, and therefore determine the wind velocity. These problems may be solved by plotting the triangle of velocities at any convenient scale, as in Fig. 14.01, although, in practice, analogue solutions have been used for the past 60 years. To the air navigator a 'computer' was the instrument used for this purpose several decades before the word obtained its modern meanings.

From the knowledge of forecast wind velocity we can determine the course to be steered in order to maintain a required track. Given a steady and accurately forecast wind, this information may be sufficient to guide the aircraft to its destination. However, significant changes in wind velocity may occur, especially during a long flight, and these will blow the aircraft from the intended track unless they are recognised and an alteration in course is made to counteract them

If the flight is in or above cloud, or over the sea, there is no method of recognising what displacements have been taking place until the navigator has another opportunity to fix position. When this can be done it becomes possible to measure the total effect of the wind during the time which has elapsed since the position of the aircraft was last determined. Using this information, and assuming that the aircraft will continue to be influenced by similar winds during the remainder of the flight, the appropriate alteration in course is made to regain the required track.

The air plot

Much of this reasoning can be done graphically by plotting every course flown as a continuous traverse on the chart. Each course is plotted as a vector, the length of which is the distance flown at the measured air speed. The result might appear as illustrated in Fig. 14.02, where the figures indicate the times at which an alteration in course was made or a fix had been obtained. The *air plot* indicates where the aircraft would have been if there had been no wind. Using distances calculated from the true air speed it is possible to locate the *air position.* If we calculate the corresponding distance from the assumed wind speed and plot a line to represent the effect of the wind from the air position, we locate the *DR, or dead-reckoning position* of the aircraft for the same instant in time. However the DR position is only an estimate which is based upon the navigator's opinion of how the wind has affected the aircraft in flight. This estimate can only be checked by fixing the position of the aircraft.

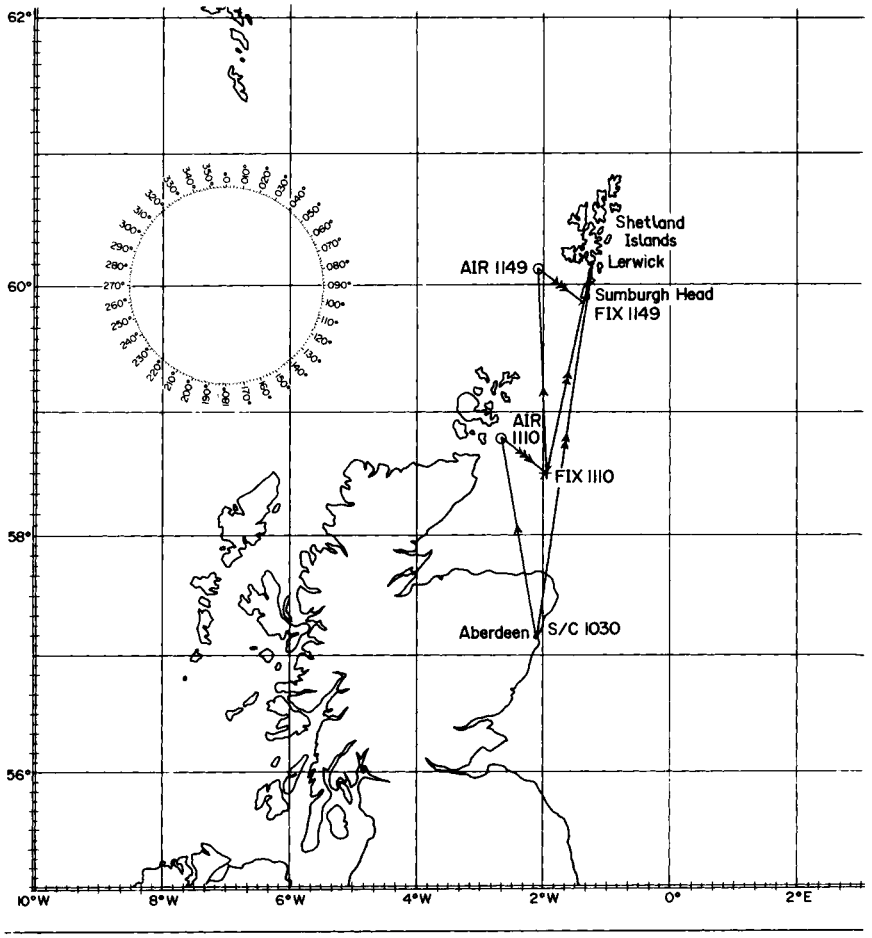
Fixing position

Fixing position may be accomplished in a variety of ways, by means of visual or radio bearing, by astronomical or electronic methods, or even by map reading when the ground is visible. The information obtained is usually in the form of bearings. Exceptions to this are fixes obtained by some electronic methods and 'pin-pointing' on a map some place on the ground which is immediately beneath the aircraft. All that we know from a single bearing is that the aircraft was located upon the *position line* represented by this bearing plotted on the chart. The intersection of two or more bearings would only fix the aircraft uniquely if all the bearings had been made simultaneously. Usually a short time elapses between observing, recording and plotting each bearing, so that the chart shows several position lines indicating locations of the aircraft at different times. The graphical method of converting several such position lines into a fix is to transfer some of them as parallel lines by amounts corresponding to the distance flown between the individual observations. The method of transfer is illustrated in Fig. 14.03.

This brief account of the methods of DR navigation indicates that the navigator has to plot straight lines on the chart to represent courses, tracks, wind vectors, bearings and position lines. Measurement of direction of all of these is important. Measurement of distance along some of them is also important.

Great circles and rhumb-lines

We must now make the important distinction between two kinds of line on the earth's surface: the great circle and the rhumb-line. We have


62°
60°
58°
56°
10°W
8°W
6°W
4°W
2°W
0°
2°E
Shetland Islands
Lerwick
AIR 1149
Sumburgh Head
FIX 1149
AIR 1110
FIX 1110
Aberdeen
S/C 1030

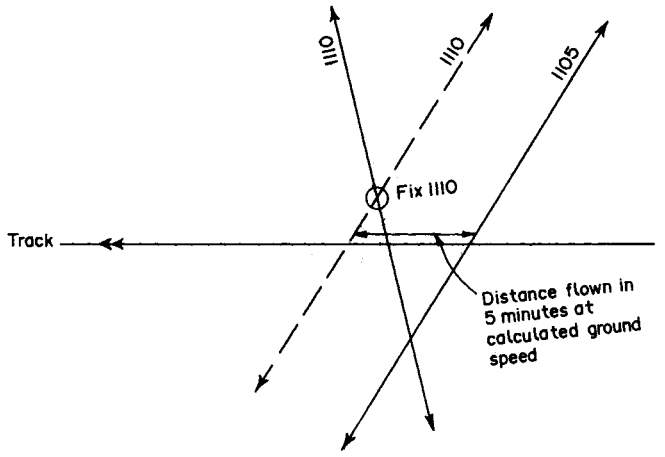

1111
1110
1105
Fix 1110
Track
Distance flown in 5 minutes at calculated ground speed

FIG. 14.02 An example of a graphical air plot for a flight from Aberdeen to Lerwick represented on Mercator's projection. The track from Aberdeen to Lerwick is 008° (true) and the distance is 182 nautical miles. This is indicated by the straight line with two arrowheads. The forecast wind velocity is 315°/50 kts and the economical cruising speed is 150 kts. The navigator therefore calculates the course to be 350° (true). This is plotted as the line with the single arrowhead, commencing at Aberdeen. The pilot sets course over Aberdeen on this heading at 1030. At 1110 the position of the aircraft is fixed in latitude 58°30′N, longitude 1°59′W, indicated on the chart by the small cross labelled 'FIX 1110'. The air position corresponding to this is labelled 'AIR 1110' and is the position that the aircraft would have reached if there had been no wind. Measurement of the direction and length of the short line joining these two points indicates the effect which the wind has had upon the aircraft during the 40 minutes which have elapsed since setting course. This shows that the total effect of the wind upon the aircraft corresponds to a wind from 340° of speed 40 kts. From the position which has been fixed to the destination is a track of 012°, and to follow this assuming the new wind velocity may be calculated as 358°. The new track and course are here plotted from 'FIX 1110', though we must emphasise that this is an approximation. The aircraft has already continued on this old course for some minutes while the navigator plotted and worked out the new wind velocity. At 1149 the aircraft crosses the coast of the Shetland Islands 2 miles north-west of Sumburgh Head. This represents 'FIX 1149' with the corresponding point 'AIR 1149' on the air plot. This confirms that the aircraft is close to the intended track, that the wind has remained constant in speed and direction. The craft is now so close to its destination that no further alteration in course is needed. Indeed it would probably have overshot Lerwick in the time needed to calculate a new course.

The need to work fast and ahead of time indicates the main difference between the practice of DR navigation at sea and in the air. The slower speed of a ship means that an alteration of course can be made from a fix even after several minutes have elapsed between the time of observation and execution of the alteration in course. The faster the speed of an aircraft the greater the distance covered in an equivalent time, and therefore increased uncertainty about position when finally altering course. This accounts for the reduction in the use of graphical methods of DR navigation in modern flying.

already defined the properties of a great circle in Chapter 3. It will be recalled that the shortest distance between any two points is the arc of the great circle passing through them. However, any great circle arc which is neither part of a meridian nor part of the equator has the property that it intersects every meridian at a different angle. This is owing to the convergence of the meridians, and the quantity γ, defined by equations

FIG. 14.03 Graphical transfer of position lines in DR navigation. Two positions resulting from observations made at 1105 and 1110 respectively are represented by the full lines. Their intersection suggests that the position of the aircraft lies to the south of the intended track. However, the position of the aircraft had changed during the 5 minutes which elapsed between the two observations. It is therefore desirable to transfer the first position line in the direction of flight by the assumed distance flown during the interval between observations. This is done by transferring the position line as a parallel straight line (shown here as a broken line). It indicates that the aircraft was north of the intended track at 1110.

(3.26) and (3.27) on p. 63, is also a measure of the total change in bearing along a great circle arc. A line of constant bearing, which is a line intersecting every meridian at the same angle, is known as a rhumb-line. This is the spiral curve on the spherical surface illustrated in Fig. 14.04.

The great circle arc has twofold significance in navigation. First, the *great circle sailing*, or track which follows the great circle arc, is the shortest distance between two places. Secondly, the path followed by a radio signal or any other kind of electromagnetic propagation or reflection between a beacon and the craft is also the arc of a great circle. So, too, is a visual line of sight. However, the comparatively low accuracy with which bearings are measured in navigation makes this distinction unemployable over the distances which visual bearings can be made. We will find, in Chapter 15, that the fact that visual lines of sight lie in the planes of great circles is important to the surveyor.

The *rhumb-line sailing*, or track which follows a rhumb-line, is important because this is the line followed by any craft which is steered on a constant heading. This procedure becomes necessary once a craft is out of sight of land and its heading must be maintained instrumentally with reference to compass or directional gyro. In short, the rhumb-line has constant direction but represents the longer distance, whereas the great circle is the shortest distance but varies continuously in direction. Clearly it is more convenient to steer a rhumb-line course if the extra distance travelled is small. Conversely a long flight or voyage may be shortened

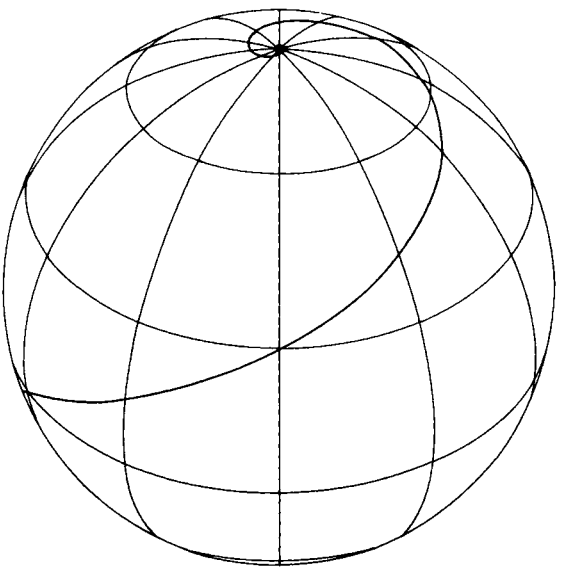


FIG. 14.04 The representation of a rhumb-line on the spherical surface.

by altering course in such a way that the great circle sailing is maintained. Thus the relationship between these kinds of line is important in practical navigation. From the point of view of choosing a projection for a chart it is valuable if either of these lines can be represented by means of straight lines. If rhumb lines are rectilinear the graphical work of DR navigation is simplified. If great circles are rectilinear the preliminary planning of a great circle sailing and the plotting of radio bearings are facilitated.

We demonstrate the differences between the great circle and rhumb-line on the earth by giving two examples. The reader who wishes to find out how these results were derived is referred to Maling (1989) or to standard works on navigation such as Admiralty (1954), Gardner and Creelman (1965) or Cotter (1966). The great circle distance from Lerwick (60°09′N, 1°09′W) to Bergen (60°24′N, 5°19′E) is 191 nautical miles (354 km). The rhumb-line distance between these places is 193 nms (358 km).* In order to maintain the great circle sailing it would be necessary to set course from Lerwick on the heading 083°(T)† and alter course at regular intervals until Bergen is approached on a heading of 088°(T). Theoretically, these changes in heading ought to be applied continuously but, in practice, they would be made in $\frac{1}{2}^\circ$ steps at intervals of 18 nautical miles. The alternative rhumb-line track requires a constant heading of 085°(T) to be maintained throughout the journey. The alterations in heading needed to follow the great circle are an inconvenience if the distance saved is less than 2 nautical miles.

The second example comprises the comparison of a trans-Atlantic flight from Halifax, Nova Scotia (44°40′N, 63°35′W) to Lerwick. The great circle distance is 2359 nms (4372 km) and the rhumb-line distance is 2452 nms (4544 km). Following the great circle therefore represents a saving of 93 nms (172 km). In order to maintain the great circle track it is necessary to alter course through more than 50°.

From these two examples we see that the difference in distance between the great circle and rhumb-line sailings varies according to the length of the arc. It can also be shown that the difference varies according to the bearing between the terminal points and their latitudes. In the limiting case of a track which coincides with a meridian there is no difference between the two lines.

*One nautical mile equals 1 minute of arc measured along a great circle. If the earth is regarded as being spherical so that the relationship between angular and linear distance is everywhere constant as in equation (1.02) then, traditionally $1' = 1$ nm $= 1853{\cdot}2$ m $= 6080$ ft. This is now known as the *Admiralty nautical mile*. The length of the *international nautical mile* is 1852·0 m.

†Tracks, courses and bearings may be defined as *true* (T), *magnetic* (M) or *grid* (G) according to which datum is used for measurement. A true course is measured from the meridian, i.e. true north; a magnetic course is measured from the direction taken by a magnetic compass, i.e. Magnetic north; a grid course is measured from grid north, as defined on p. 33.

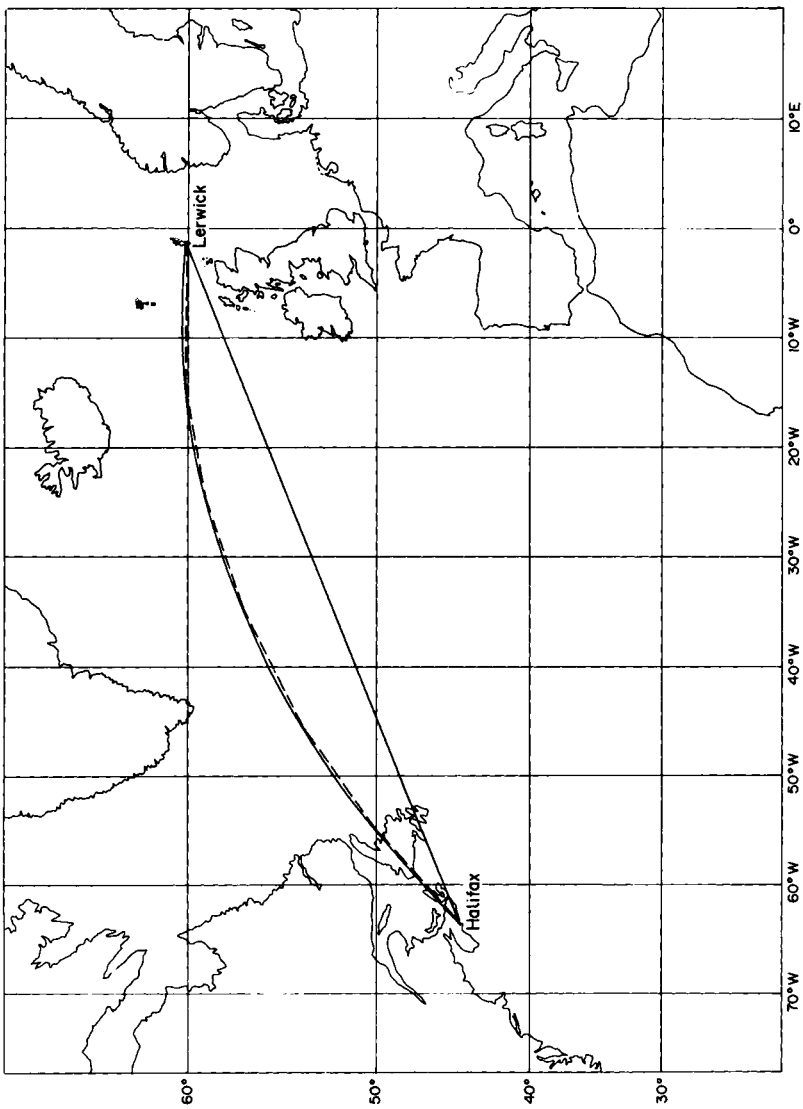


FIG. 14.05 Part of the North Atlantic Ocean on Mercator's projection showing the great circle (curved) and rhumb-line (straight) tracks between Halifax and Lerwick. A composite track approximating to the great circle is indicated by the broken line. This is composed of a series of shorter rhumb-lines which approximate to the curve of the great circle.

The usual navigation practice is to follow rhumb-line tracks unless the voyage or flight is so long that the great circle sailing offers an appreciable saving of distance, and therefore of time and fuel. Even then the navigator uses a *composite track* (Fig. 14.05) which divides the great circle into a series of shorter rhumb-line elements. This means that the craft is steered on a succession of constant headings, altering course at suitable intervals to keep close to the great circle.

Suitable projections for navigation charts

The various requirements for the projection of a navigation chart can now be summarised as follows:

- Since angular measurement is an important feature of DR navigation, a *conformal projection is obligatory*. There are plenty of these to choose from, though usually the choice is restricted to the conformal members of cylindrical, conical and azimuthal classes. These are respectively, *Mercator's projection*, *Lambert Conformal Conical projection* and the *Stereographic projection*. Moreover, these are generally used in their normal aspects.
- Since plotting and measuring distances is an important aspect of chart use, a projection in which linear distances are truly represented would seem to be desirable. However, we have already explained in Chapters 5 and 6 that it is impossible to create a plane map in which the principal scale is preserved at all points and in all directions. The best that we can hope to do is to use a projection in which the particular scales do not change too rapidly from place to place on the chart.
- Since craft are steered along rhumb-lines *a projection which shows rhumb-lines by means of straight lines is valuable.*
- Similarly, *the representation of great circle arcs by means of straight lines is desirable.*

Clearly the last two requirements are incompatible with one another. We have demonstrated the difference on the earth between the great circle and the rhumb-line joining two points. We cannot, therefore, expect the same projection to depict both as straight lines, for this would be the same line and would violate the fundamental principle of one-to-one correspondence between points on the earth and the chart.

There are two conformal projections which satisfy one of these additional requirements. Mercator's projection (pp. 209–217) has the important additional property that *all rhumb-lines are straight lines*. This follows directly from the fundamental property of all normal aspect cylindrical projections that the meridians are represented by parallel straight lines, and from the special property that Mercator's projection

is conformal. It follows that any straight line drawn across a Mercator chart intersects every meridian at the same angle. Since there is no angular deformation, this straight line satisfies the definition of a rhumb-line as being a line of constant bearing.

The Lambert Conformal Conical projection has the additional advantage that great circle arcs are almost rectilinear. We must emphasise that this is not strictly true. For example, this assumption is not acceptable to a surveyor using this projection as the base for topographical mapping. However, within the limits of most aeronautical charts prepared on this projection, which are smaller than 1/500 000 scale, and within the margins of error which have to be accepted in steering a craft and taking bearings from it, the departure of the great circle arc from a straight line is small enough to be ignored. Since the projection is also conformal it is possible to plot radio bearings as straight lines.

Only one projection strictly satisfies the property that all great circles are represented by straight lines. This is the *Gnomonic projection* (Appendix I, No. 7), which is a member of the azimuthal class. However, this projection is not conformal and it has the additional disadvantage that the radial particular scale ($\mu_1 = \sec^2 z$) increases rapidly from the point of zero distortion. This in turn means that measurement of distance on the Gnomonic projection is also unreliable unless special measures are taken to correct for the rapid changes in particular scales. It can be done, but it is not convenient for use in practical navigation.

Projections for nautical charts

In marine navigation, where graphical plotting is still a normal practice, the role of Mercator's projection is unassailable. The practical reasons for its popularity have been admirably summarised by Stigant (1947), as follows:

> Among sailors there is a kind of divine belief in the Mercator. It has two properties which fit needs absolutely and precisely.... They are... the fact that a straight line drawn on a chart is a line of constant bearing, and the other not less important property is the parallelism of the meridians and parallels which permits you to put the compass rose at one end of the chart and, if your parallel ruler is long enough, to transfer a line of bearing to the other end. These advantages figure very much in the plotting techniques used by the average navigator, who is not necessarily a cartographical expert. He is used to stepping off a distance from the latitude graduation of the Mercator and transferring a bearing from one end to the other merely by the use of his parallel rule. We cannot... abrogate these advantages lightly or without being pretty sure that the reasons are sufficient.

It follows that if rhumb-lines are represented by straight lines, great circles must be curves. On Mercator's projection great circles are represented by curves which are convex towards the nearer pole, as illustrated in Fig. 14.06(a). If a decision has been taken to follow the great circle

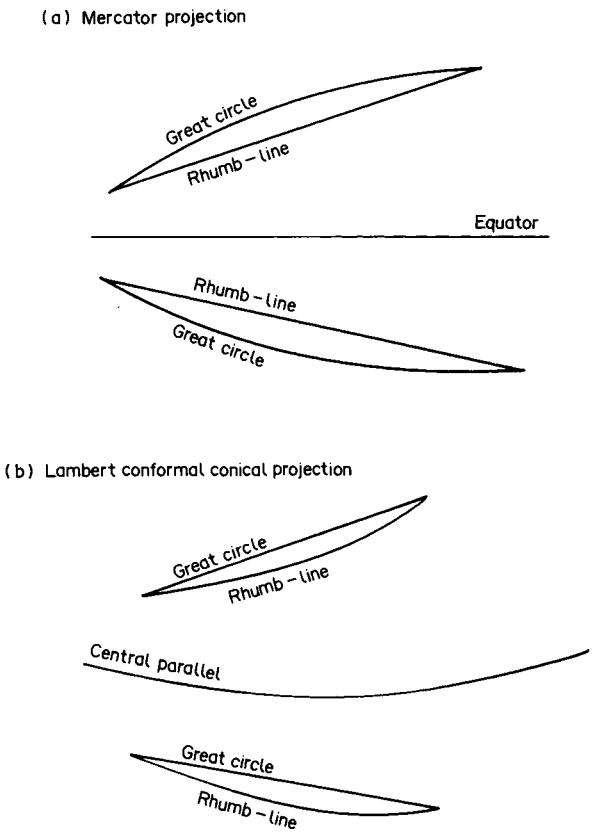


FIG. 14.06 The representation of great circles and rhumb-lines on (a) Mercator's projection and (b) the Lambert Conformal Conical projection. Note that on the Lambert Conformal Conical projection great circles are not depicted by perfectly straight lines as illustrated here, but this assumption is made for all practical purposes in navigation.

route between two places, a curve corresponding to this track must be plotted on the charts to be used. This can be done by computing the geographical coordinates of certain points on the great circle, for example to find the latitude where the arc cuts certain meridians. Alternatively, the route might be plotted first on a Gnomonic chart and these intersections transferred to the Mercator charts. Normally this kind of work is done when planning a voyage or flight and is completed well in advance of putting to sea or take-off. Therefore the slow job of plotting the great circle curve does not normally figure in operational navigation.

However, radio bearings may still be used to fix position and the need to plot a great cricle bearing as a position line needs a rapid method of locating part of the curved line using the simple routines of DR plotting

and a method which does not occupy too much time. The navigator uses a *conversion angle* which is the arc-to-chord relationship between the great circle arc and the rhumb-line chord as these appear on the Mercator chart, e.g. Figs 14.05 and 14.06, measured both at the radio beacon and at the assumed position of the craft. For a spherical earth this angle can be shown to be $\gamma/2$ where γ is the convergency defined in Chapter 3, p. 62. Application of the conversion angle at each end of the arc, as shown in Fig. 14.08, p. 306, allows the navigator to plot a short straight line in the vicinity of the craft's DR position to represent the position line corresponding to that part of the great circle arc.

The Mercator projection has the important disadvantage that accurate linear measurement is difficult. This results from the increase in particular scale as a function of $\sec\varphi$ (10.65), p. 213. Thus a distance of 5 cm on the edge of the chart in a higher latitude represents a shorter distance than 5 cm on the edge of the chart near the equator. The change in scale is particularly rapid north of 60°N (or south of 60°S). It follows that a reliable measurement cannot be made using a ruler with equidistant subdivisions and converting things to distance on the earth through the representative fraction (principal scale) of the chart. Measurement on Mercator's projection has to be made with dividers, comparing the separation of the points with the latitude subdivisions along the border of the chart, or along one of the meridians which have been closely subdivided for this purpose. Figure 14.07 illustrates how this is done. The comparison must always be made in the same latitude as the line to be measured, setting the dividers along the border symmetrically about the mean latitude of the line. In this way the variations in particular scale tend to be compensated, but the measurement is an approximation nevertheless. Since latitude subdivisions are in minutes of arc, or multiples thereof, and since 1′ of latitude measured along a meridian corresponds to 1 nautical mile, comparison of the dividers with the meridional border of the chart gives the distance in nautical miles. This technique emphasises further why navigators prefer to work in units of nautical miles and knots. This in turn explains why there is no special reason for the scale of a nautical chart to have some integer value such as 1/500 000 or 1/1 000 000, but may be 1/545 000 or 1/997 562. It should be noted that clumsy representative fractions such as the last pair are no longer used in the preparation of new charts. Consequently they are gradually disappearing as new (in Britain the metric) charts are produced. However, the replacement of chart cover for the world's oceans is a long, slow business, and until that is accomplished there will still be a few charts made in times when a major aim was to produce the largest-scale chart which would fit a standard size sheet of paper. There is, of course, no real objection to the use of clumsy representative fractions. The marine navigator probably has to make more measurements of distance than any other user of

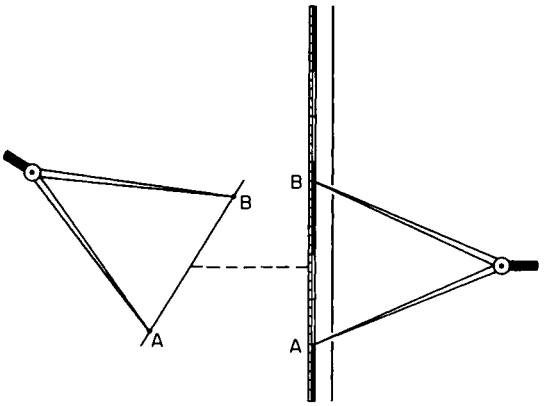


FIG. 14.07 Measurements by dividers on Mercator's projection. This is always done by comparing the spread of the dividers against the latitude subdivisions along the eastern or western edge of the chart in the same range of latitude as the line to be measured.

maps or charts, but never makes the kind of scale conversion typical for use with topographical maps.

Projections for large-scale nautical charts

Although we have dismissed the Gnomonic projection as having only limited use for route planning, its name appears again in marine charting, for the words *'Gnomonic projection'* appear on all Admiralty charts of scale 1/50 000 and larger. These are the charts of port and harbour approaches or navigable rivers. At this scale the chart does not extend more than about 15 km from the centre of the sheet. If, therefore, the centre of the chart represents the origin of the projection, the linear distortion and angular deformation at the edges are less than the zero dimension and too small to be measured. This confirms the conclusion reached in Chapter 11, pp. 221–223, that the choice of projection for a large-scale map or chart is often unimportant. Perhaps this is just as well, for the description 'Gnomonic projection' on large-scale charts is incorrect. The Hydrographic Department admits as much in Admiralty (1965), stating that the projection is really a version of the Polyconic projection. The equations used to obtain the coordinates of graticule intersections are those of the Polyconic projection; but these intersections are joined by straight lines to represent the 2′ or 4′ parallels and meridians which appear on the chart. Strictly speaking this converts a Polyconic projection of Group A into a version of the *Polyhedric projection*, which is a pseudocylindrical projection of Group C. Of course the distinction is trivial in conventional map use for it cannot be measured

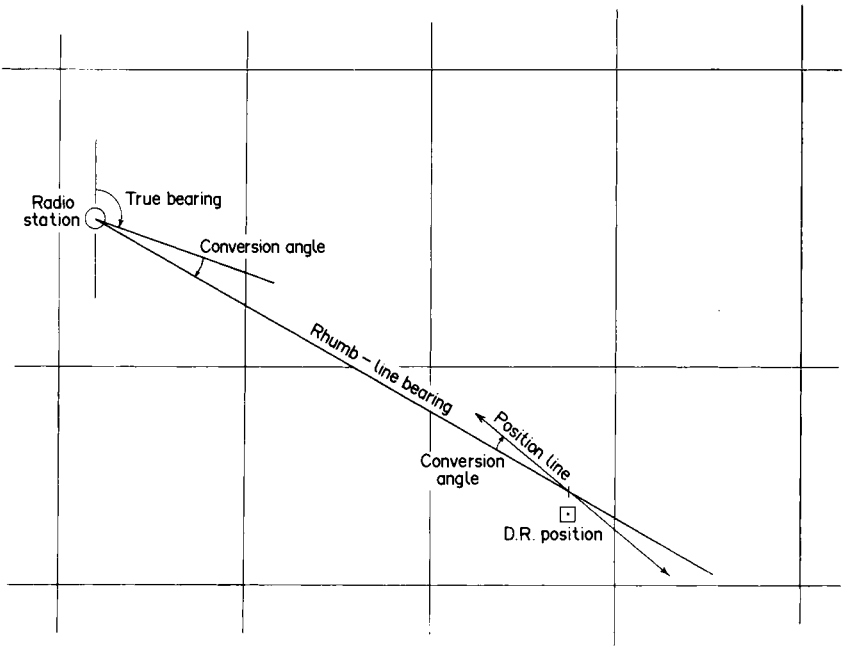


FIG. 14.08 The representation of great circle bearings as position lines on Mercator's projection by use of the conversion angle. The path of a radio signal between a transmitter and a distant craft follows the great circle arc passing through them. Therefore the measured direction of a radio signal refers to the bearing of the great circle, not the direction of the rhumb-line joining them. In order to plot an observed bearing as a rhumb-line on Mercator's projection, and also plot the direction of the position line as part of the great circle arc, it is necessary to determine the conversion angle by calculation or from a nomogram. This angle is applied to the great circle bearing measured at the radio station to find the rhumb-line to be plotted on the chart. In order to orientate the position line to the great circle it is necessary to apply the same conversion angle in the opposite sense to the rhumb-line plotted near the DR position of the craft. It follows that the position line plotted on he chart now represents a short element of the great circle arc.

on a chart, but it is useful to set the matter straight on this matter and use the correct name for the projection. One question remains: why is it called a Gnomonic projection? We can only assume that most navigators know about the special property of the Gnomonic projection, and since entry to a port or harbour is normally made by visual alignment with buoys or shore features, the fact that visual bearings may be presumed to be straight lines is of some comfort to them.

Major technological advances have affected marine navigation just as they have changed other branches of making and using maps and charts. Thus the introduction of various systems of navigation satellite, which

were briefly mentioned in Chapter 1 for their applications in modern geodesy, have made position fixing possible anywhere in the world with orders of accuracy which are superior to any of the navigation aids dating from the mid-twentieth century. These employ the principles of the doppler navigator to measure distances from the craft to a group of artificial satellites whose orbital positions are precisely known for any given instant of time. The methods of inertial navigation preserve a record of position by measuring and storing information about the accelerations of the craft in three dimensions. The methods of GIS applied to navigation requirements go far towards the production of the *electronic chart*. At the end of the 1980s, however, there are still some vessels which do not yet carry any sophisticated aids and, in any case, the navigator must be prepared and able to back up any failure in equipment or power supply by using the traditional methods of DR navigation on a paper chart.

Projections for aeronautical charts

In the early days of flying, before 1939, an airspeed greater than 150 mph was the exception. Therefore it was possible to use graphical methods of DR navigation, and the methods of air navigation were very like those used at sea. The evolution of air navigation from marine navigation was reflected in the design of the aeronautical plotting charts of that time, which were based on Mercator's projection (and even still showed the depth of the sea). Throughout World War II the Mercator plotting chart was used by the Royal Air Force, almost to the exclusion of all other types, but as flying speeds increased it became more and more difficult for the navigator to maintain an up-to-date plot graphically. Consequently analogue instruments were developed to do this part of work. For example the *air position indicator*, being an instrument accepting input from both the airspeed indicator and the gyrocompass, was used to maintain a continuous record of the air position, which otherwise had to be plotted on the chart with protractor and dividers.

Associated with the increased performance of postwar aircraft and the increased density of traffic came the need to extend and improve the network of radio aids allowing the navigator to fix position or to home to a beacon. Consequently much airline navigation, following regular routes to a fixed timetable, may be categorised as *operating along great circle tramlines* by Anderson (1968). In addition, various methods of electronic distance measurement have been applied to navigation in the form of the *Rebecca-Eureka*, *DME*, *Vortac* and *TACAN* systems. These locate the position of an aircraft by bearing and distance from the beacon. Thus the requirements for aeronautical charts have changed from the need for a document upon which DR navigation can be plotted to a chart on which the great circle tramlines are rectilinear and distances are easy

to measure. This has led to the replacement of the Mercator chart by those based upon the Lambert Conformal Conical projection as the base for modern aeronautical charts. Other arguments favouring this change are to be found in Freer and Irwin (1951) and Peake (1947). The advantages of showing great circle arcs by straight lines have already been stressed. By using versions of the Lambert Conformal Conical projection with two standard parallels, the particular scales do not change rapidly within the limits of the single chart and therefore distance can be measured with a ruler. The projection had the disadvantage that rhumb-lines are curved, as shown in Fig. 14.06(b), but if there is no need to maintain a graphical air plot this defect is unimportant. A futher disadvantage is that measurement of bearings, courses and tracks must be made using a separate protractor which must be orientated to the meridian through the point where an angle is to be measured.

Since the meridians converge there is no possibility of using a printed compass rose, as on the Mercator chart, and using a parallel ruler to transfer lines to any place on the chart.

After the paper chart

Today the aeronautical chart printed on paper is being replaced by other kinds of graphic display. In the first edition of this book, which was written in the early 1970s the author described one version of aeronautical chart which was already being superseded. This was a map mounted on rollers, the movements of which were controlled by monitoring of track and ground speed by doppler equipment to indicate the position of the aircraft. This kind of configuration normally requires a special map prepared on an oblique aspect Mercator projection in which the line of zero distortion is the intended track of the aircraft and was consequently only suitable for use aboard aircraft which were repeating the same flight to the exclusion of all other activities. Even then this analogue system was being replaced by the use of charts reproduced on film which could be projected to a screen and linked to the avionics system so that the chart image moved automatically to maintain the aircraft's position at the centre of the screen. Moreover, the scale of the projected image could also be controlled by the pilot or navigator to meet their immediate needs. Some of the methods used to prepare these have been described by Honick (1967). Today, of course, these rather clumsy analogue systems have been replaced by sophisticated computer graphics which provide similar, continuously changing displays on the screen of a visual display unit.

Polar navigation

One aspect of air navigation which hardly affects the mariner is that of navigation near the geographical poles. Special techniques of high-lati-

tude navigation were developed in the years immediately following World War II when the possibilities of operating commercial flights on trans-polar routes became likely. In high latitudes definition of true direction is a major problem because the meridians converge to a point. During the past 40 years direction on aeronautical charts has been referred to the *Greenwich Grid*, a plane cartesian grid with one axis coincident with the Greenwich Meridian. This grid provides a constant datum to which courses, tracks, bearings and magnetic variation may be referred, greatly simplifying the graphical work. The technique of using the Greenwich Grid in high latitudes has been summarised by Beresford (1953) and Hagger (1950). Traditionally the projection suitable for conformal representation of the polar regions is the normal aspect Stereographic projection, but later years the *Transverse Mercator projection* has also been used for polar navigation. The *USAF Global Navigation and Planning Chart, GNC-1*, of scale 1/5 000 000, is based upon this projection and the methods of using it have been described by Dyer (1971).

CHAPTER 15

Surveying and map projections

> What we really want is a system which will enable us to consider the earth as flat over as wide an area as possible and for as many purposes as possible, and so avoid troublesome curvature corrections to our observations in the vast mass of minor surveying which does not need to split seconds...
>
> There must come a time, however, when we can no longer neglect such corrections, whether because our purpose requires more accuracy or because we happen to be working a long way from the error-free line of the projection. In such cases we must apply corrections, but we want them to be simple and rapid...
>
> M. Hotine, *Conference of Commonwealth Survey Officers, 1947, Report of Proceedings*

Introduction

In the next two chapters we are concerned with the definition of planimetric positions of survey control points. Ultimately it is the accuracy of their location which determines the accuracy of the entire map; therefore the relationship of observations made on or near the geoid to their coordinates referred to a plane grid is important. Plotting of map detail is generally done by photogrammetric methods. The effect of earth curvature upon this stage of the work is investigated in Chapter 17.

We have already seen in Chapter 1 that the results of surveys are computed at the natural scale of 1/1. It follows that distortions and deformations which are too small to be detected on maps at scales of 1/250 000, 1/25 000 or even 1/2500 are measurable quantities on the ground. Consequently the surveyor must also be concerned with the effects of earth curvature, and must apply suitable corrections to observed or computed angles and distances in order to locate planimetric position on the required projection. However, circumstances arise when the accuracy requirements for a small job permit some relaxation in the choice of reference surface and projection, even to the extent of being able to regard the earth as being flat over the whole area to be mapped. Such assumptions depend upon the purpose of the survey, which, in turn, determines the precision of the instruments used and the observing routines to be adopted. If the deformation resulting from earth curvature is

less than the accuracy of measurement of angles and distances, errors in position attributable to choice of reference surface and projection are too small to be considered. In other words, the concept of the zero dimension still applies, although it is of different magnitude to that applying to maps and charts. It is therefore desirable to introduce this specialised application of the use of projections by addressing ourselves to some consideration of the application of projections to surveying and mapping.

Projections in actual use for control surveys, cadastral surveys and topographical mapping

It was stated in Chapter 6, p. 107, that the special property of conformality is a necessary requirement for large-scale and topographical maps. This has not always been so. Many national surveys were originally based upon projections which are not conformal. The use of Cassini's projection in Britain, various versions of the polyhedric projection in central Europe and the Polyconic projection in the USA are typical examples. At the smaller topographical scales, equal-area projections such as Bonne's were often used. Hinks (1921) has described how the need for conformal topographical maps arose primarily as an artillery requirement during World War I. Since that time the majority of national surveys have been converted to conformal projections and all modern large-scale and topographical map series are now based upon them.

In an evaluation of world mapping in 1980, carried out for the United Nations by Brandenberger and Gosh (1985), 27 different named projections were listed as still being in use for topographical, cadastral and engineering surveys. However, the statistics presented by them need rather careful interpretation because their list is encyclopaedic and each minor variation in name warrants a separate entry. For example, they refer to five different versions of the Transverse Mercator projection, although there are only trivial differences between some of them; the Lambert Conformal Conical and Lambert Conical Orthomorphic are listed separately, although this is only a variation in name. Moreover many countries use several projections for different purposes but each application has been entered in this list without distinction. Table 15.01 is a heavily edited version of their list. If we adopt this more realistic interpretation we find that only three systems of projection are now of real importance in terms of the land area covered. These are:

- Transverse Mercator projection, 85%
- Polyconic projection, 10%
- Lambert conformal conical projection 5%.

One variety of the first, known as the *Universal Transverse Mercator* (*UTM*) system, was originally introduced by the United States Army in

TABLE 15.01 *Projection systems in use in the world in 1980. Areas in thousands of km². Source: Brandenberger and Gosh (1985)*

Projection System	Africa	N. America	S. America	Europe	Asia	USSR	Australia and Oceania	World
UTM	25 583	15 950	12 055	1 164	18 081		517	74 071
TM	5 852	1 907	3 564	1 596	13 480	22 402	291	49 093
Polyconic	1 784	9 372	2 789		2 896		218	16 842
Gauss–Krüger			4 673	1 002	9 969			156
Lambert Conformal Conical versions	602	2 258	769	1 029	3 470			8 129
ACT Grid							7 687	7 687
Oblique Mercator versions	590			41	336			966
Bonne	164			193	569			926
Cartesian			1 139					1 139
Polyhedric			1 285	340	372			1 997
Stereographic			163	134	10			307
Hatt (Azimuthal equidistant)				132				132
Cassini–Soldner		5		70			21	96
Others	7	1						8

the early 1950s as a uniform projection system for military mapping. Although its introduction initially met with some opposition, its use as the projection base for both civilian and military purposes is now well established in the majority of Western countries and much of the Third World. Figure 15.01 shows the projection systems in current use, and demonstrates the great importance of the UTM.

Reference has already been made to the use of more than one system for different purposes in the same country. In some countries the projections used for cadastral purposes are still a mixture of Conformal Conical, Polyconic, Cassini and Azimuthal Equidistant systems, with many local variations of origin and orientation, which are relics of the nineteenth century. The existence of extensive registers of legal titles to land, defined in terms of coordinates based upon these systems, fosters their preservation. Consequently the results of local surveys must often be computed and recorded on projections which differ from that used by the national survey. Of particular importance is the stress which has been laid in some countries upon the accuracy requirements for cadastral surveys to the extent that the projections used for topographical mapping have sometimes been unacceptable for cadastral work. We return to this subject later.

Another example of duplication occurs with projections intended for civil engineering use. For example, in the United States the individual State or local coordinate systems in use are quite different from those employed by the federal mapping agencies. Until the 1950s the US Geological Survey employed the Polyconic projection as the base for national mapping, creating the characteristic maps comprising graticule quadrangles of size $7\frac{1}{2}' \times 7\frac{1}{2}'$ and $15' \times 15'$ for the 1/24 000 and 1/62 500 scales respectively. The positions of control points referred to such maps have to be expressed in geographical coordinates and are therefore awkward to use in everyday survey practice. Consequently the State Plane Coordinate System was introduced in 1936, primarily on the initiative of highway engineers, but of considerable importance for all kinds of engineering surveys. The individual projections forming the State Coordinate systems were based upon the Clarke 1866 ellipsoid and what is now known as NAD27, the North American Datum defined in 1927. The separate projections or groups of projections were established for each State according to the general principle that the scale error due to the projection should never exceed 1/10 000. States with a predominantly east–west extent employ the Lambert Conformal Conical projection, and those with greater north–south extent are mapped on the Transverse Mercator projection. Several States have more than one Transverse Mercator zone, and a few States have a combination of both Lambert and Transverse Mercator. A small part of Alaska makes use of an oblique Mercator projection. A certain amount of revision of the system has been required

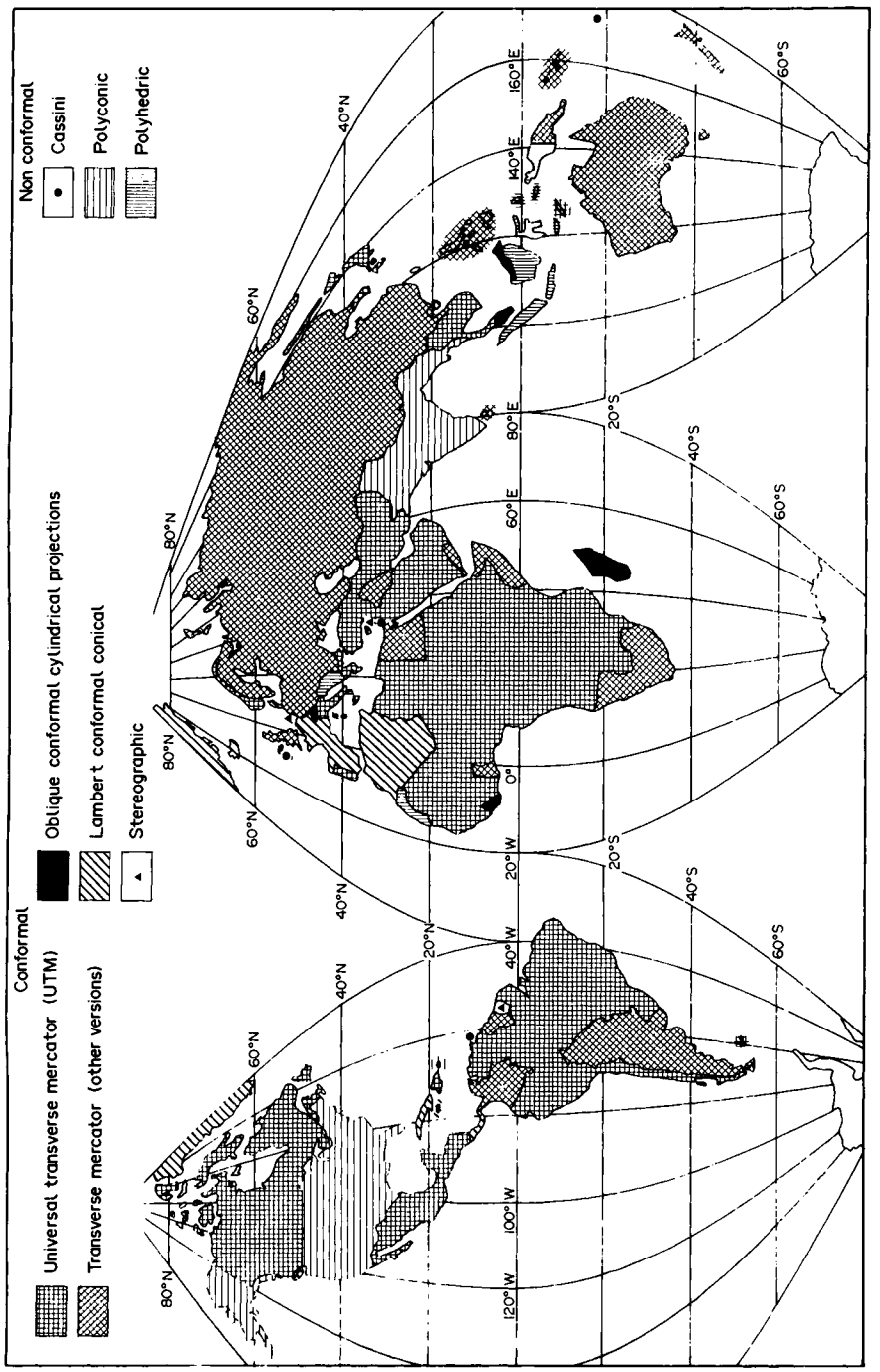


FIG. 15.01 World map indicating the projections currently used for large-scale mapping. The primary sources for information are *World Cartography* X, 1970; XIV, 1976; XVII, 1983, which give information about the projections used for basic scale mapping by national surveying departments. Because of the continuing change to some version of the Transverse Mercator projection, many countries which are here shown as using another projection are also using the Transverse Marcator projection for the smaller-scale topographical map series, or are gradually replacing existing basic mapping by a new series based on the Transverse Mercator.
The projection used for this world map is a recentred version of the Sinusoidal (No. 30 in Appendix I) comprising five lobes, with central meridians in longitudes 100°W, 70°W, 20°E, 30°E, and 130°E.

with the conversion of the US geodetic control network from NAD27 to the new NAD83 North American Datum. For a comprehensive list of the changes see Snyder (1987a).

Published maps frequently have a long life before they are finally revised and replaced. Consequently the use of an older projection system may linger on in map use long after the policy decision has been taken to change to a conformal and national system of projection, and after the national trigonometric network has been recomputed on the new system. For example the policy decision to change Ordnance Survey maps to the National Grid was made in 1939 (p. 335) and accepted by Parliament as government policy in 1945. The replacement of the basic 1/2500 mapping commenced at that time, but it took until the middle 1980s to complete this. Similarly the USGS began to convert all topographical mapping to UTM in the 1950s, but the number of sheets to be changed is so great that many decades will elapse before the last of the quadrangles based upon the Polyconic projection has been replaced.

The nature of survey methods

In order to locate the points by traditional survey methods various combinations of angular and linear measurements are used. In a study of the influence of earth curvature upon the results of control survey, it is necessary to emphasise the following fundamental geometrical concepts relating to the angular measurements:

Angular measurements

- Angular measurements are either horizontal or vertical; these directions being established with respect to an horizontal plane, as defined by a spirit bubble, or to a vertical axis, as defined by the cord of a plumb-line suspended from the instrument.
- The point from which the observations have been made is a point in space. Therefore it must always be assumed that the location or relocation of an instrument over this point in order to measure distances after measuring angles, etc., is always accomplished without error so that all such information is referred to the same geometrical point.
- The accuracy with which angular measurements can be obtained may be better than 1 second of arc. There are a few theodolites intended for geodetic work which can be read directly to 0″·2; a number of different models of theodolite can be read directly to the nearest 1″ and there are many lower-order instruments which read to the nearest 20″ or 30″. The precision of the resulting angles may be improved by

repetition of the measurements using standard observation techniques to avoid the introduction of any systematic errors.

- All observations are made along the shortest path between the instrument and the point observed. Because the arc of the geodesic on the spheroid or the great circle on the sphere is the shortest distance between two points it follows that field observations are made along geodesics and not along straight lines. This is precisely the same phenomenon as radio direction-finding signals following great circle arcs rather than rhumb-line chords, which was described in Chapter 14.

The size of the conversion angle between the arc and chord depends upon the direction and length of the line. For the use of radio bearings in navigation this may amount to several degrees, but it is very much smaller when applied to visible lines of sight, which are seldom much more than 100 km in overall length. For example, the greatest arc–chord correction which had to be applied to the lines observed during the primary retriangulation of Britain was for the 116 km line from Healaval (Barra) to St Kilda (OS, 1967) and amounted to an angle a little more than 1′30″. However at 116 km this corresponds to a linear displacement on the ground of 52 m. An error of this magnitude is too large to ignore, so that it is necessary to apply the arc to chord conversion for much shorter lines than this extreme example.

Linear measurements

Throughout the history of surveying up to the 1950s it was far easier to obtain reliable angular measurements than those of distance. Then *electromagnetic distance measurement* (EDM) replaced the tape or chain, and it was both faster and more accurate. The accuracy of EDM techniques is normally expressed in units of *parts per million*, thereby emphasising the huge improvement over taping or optical methods of measuring distance, which seldom exceeded a relative accuracy of 1/5000 even under ideal conditions.

It is necessary to apply two corrections to the measured distances irrespective of how these have been obtained:

- The required length is the *horizontal distance* between two points, whereas the observed distance is the *slope distance*. This means that a correction must be applied which requires the additional information of the observed angle of elevation (or depression) between the stations or calculated from the measured differences in height between them. The correction is simply the solution of the right-

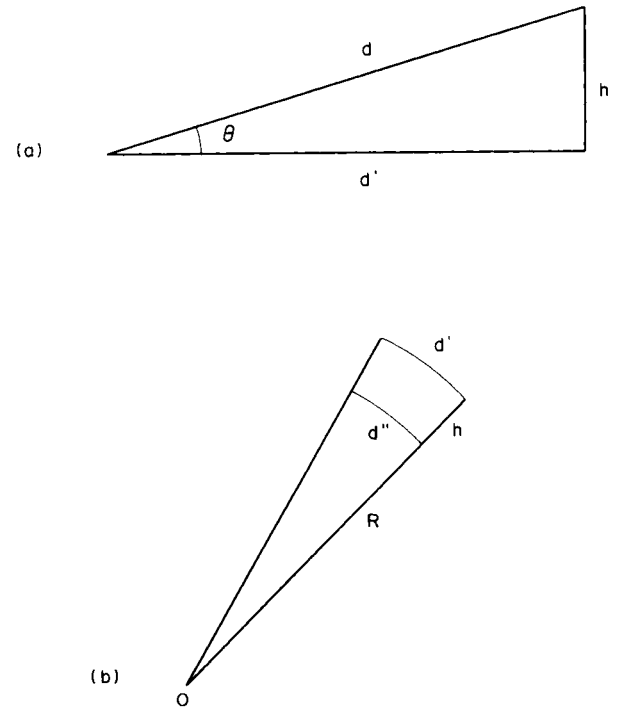


FIG. 15.02 The corrections to be applied to linear measurements, (a) for ground slope, θ, and (b) for height above (or below) the geoid (or the reference spheroid).

angled triangle in Fig. 15.02(a) or

$$d' = d \cdot \cos\theta \qquad (15.01)$$

where d is the slope length and θ is the angle of elevation between the stations.

The second correction is that relating linear measurements to the surface of the spheroid, by making an allowance in proportion to the height of the surveyed line about the reference figure. Often the work is of such low accuracy, or the ground height is so close to mean sea-level, that this correction may be ignored. Where it has to be applied it is sufficient to make the simple proportional correction for height indicated in Fig. 15.02(b)

$$d'' = d' \cdot [R/(R+h)] \qquad (15.02)$$

where d′ is the distance corrected for slope, and h is the height of the line above the spherical surface of radius R. For example a distance of 1000 m at an altitude of 1500 m is equivalent to a distance of 999·76 m at mean sea-level, because of the curvature of the earth.

Although it may be difficult to justify the application of such a small correction to tape measurements it is well within the capabilities of EDM to measure differences in length of this order of magnitude. Dale (1976) has emphasised the need for this correction in places such as the highlands of East Africa, and Vincenty (1989) upon those to be applied on the High Plains of the middle west States of the USA.

The introduction of photogrammetry and EDM heralded two major technological advances in surveying practice which have now become commonplace. A third, and more far-reaching, revolution has come about with the introduction of satellite-based and inertial methods of fixing position. It is important because it largely eliminates the traditional techniques of fixing position. Fixing position by means of doppler measurements to and from artificial satellites, which are the principles of Navsat and GPS, together with the inertial methods of fixing position by *inertial surveying systems* (*ISS*) are essentially 'black-box' methods of surveying. For a summary of the instruments and methods see various review articles published during the past decade, for example Cross (1986). These require little surveying skill; merely an ability to press a button and read the displayed coordinates of the position of the instrument. Moreover, the need for maintaining continuity between observations has now largely disappeared. In conventional surveying it has always been desirable to carry the control through an unsurveyed area until it can join up with other survey control. Thus a traverse, or even a humble line of levels, must start from a point of known position or height and *close* to another point of known position or a bench-mark of known height, and this may require many additional observations, for the methods of survey adjustment usually depend upon making such connection between existing survey controls. Sometimes delays of months, or even years, have occurred before the opportunity can be seized to observe the missing connections between surveys which are otherwise complete. Using GPS or ISS it is sufficient to transport the equipment to a site and fix the position of the instrument. It is no longer necessary to ensure that other survey stations are either visible or accessible on foot. Nevertheless, it is still too early to state that the use of GPS and other systems has wholly replaced conventional surveying methods. There are many applications for which the black-box methods still appear to be too expensive; but the same objections were raised a generation ago concerning the use of both aerial photography and EDM in the early days of their availability. However, the branch of engineering surveying known as *setting out*, which involves locating and marking points on the ground where roads, bridges, pipelines and other major structures are to be built, will still need to be based upon ground observations of the

traditional kind even after the alternative methods of fixing position have become widespread.

Projection calculations

In relating position on the curved surface of the earth to that on a plane, some of the computations which need to be made are additional to, and may even replace, the computations used in conventional cartography where the main job in plotting a new map has been to compute and plot the master grid coordinates of the new graticule as described in Chapter 8. In the computation and plotting of control survey it is usually required to determine the position of a point within the projection system with reference to other points, using the observed bearings and distances to fix the new station. For much large- and medium-scale work it is usually unnecessary to transform from geographical coordinates to map co-ordinates; indeed geographical coordinates of point within the survey scheme need never be known. Instead the computations and plotting are carried out from point to point working in grid coordinates only.

However, it is impossible to ignore altogether the relationship between geographical coordinates and grid coordinates. From the theoretical point of view it is more satisfactory to develop the customary equations relating geographical to grid coordinates as a preliminary to the derivation of those used for other purposes. This, indeed, originates from the need to use projection tables for the computations, particularly those for conformal projections of the spheroid in which we shall see there are numerous small terms to be introduced. These may have only a small effect upon the results, but they are awkward to calculate. The projection tables were logically based upon the relationship of different quantities to the geographical coordinates on the spheroid and each computation was based upon tabulated values derived in the same fashion. There is no reason, apart from that old problem of ease of computation, why terrestrial positions should not have been defined in three-dimensional cartesian coordinates.

From the practical point of view the geographical coordinates create a standard reference system so that, if it is required to transform positions of points from one projection system to another, a convenient method is to convert from the plane coordinates of one projection back into geographical coordinates on the earth and then into the coordinates for the new projection. This is described in Chapter 19. The utility of the method suggests that the starting point for the calculations may be either geographical coordinates or grid coordinates. Thus, starting from knowledge of the geographical coordinates of points we may determine the grid coordinates for a particular projection, as described in Chapter 10. We shall refer to this in future as the

geographicals to grid transformation.

The inverse process is, of course, to transform input grid coordinates into geographical coordinates and is therefore known as the

grid to geographicals transformation.

The commonest procedure in conventional survey computation is to work from point to point from observed or calculated bearings and distances relating a *known point* (which has already been located within the grid) to an *unknown point* which has not yet been fixed. This may be done by applying suitable corrections to angles, distances or coordinates which take into consideration the special properties of the projection in use. The nature of the corrections which may be applied comprise:

- Determination of the *convergence* at a point within the projection. This is the angle at a particular point, A' in Figs 15.05 and 15.06, made between the direction of grid north and true north or the angle between the grid Northing line and the meridian at the point. We shall see that this angle C, or γ, has to be calculated in order to determine the true bearing, or azimuth of a line. It is necessary to be able to calculate convergence either from input (φ_a, λ_a) geographical coordinates or from input (E_a, N_a) for the corresponding points A or A'.
- Determination of the *scale-factor* at a point within the projection. This is the particular scale at a point A'; the difference in terminology arising simply from the fact that scale-factor is the name normally used for the parameter in surveying and topographical cartography. It is possible to determine the scale factor from either (φ_a, λ_a) or from (E_a, N_a), although it is rare to have to determine the scale-factor from geographical coordinates.
- Determination of the *Arc-to-chord* or *(t–T) Correction* to an observed bearing, which has already been mentioned. It is necessary to determine this for a pair of points whose grid coordinates are known.
- Correction to determine spheroidal distance from plane grid distance. The scale-factor described above applies to one point only. All other points have different scale-factors. Therefore to apply a correction to the measured or computed length of a line, it is necessary to form an integral of the particular scales at all points making up the line. The usual way of doing this is to apply *Simpson's Rule* of numerical analysis, as described on p. 448.

Assumptions about the earth's figure

The need to employ a reference figure having known geometrical characteristics was introduced in Chapter 1. We have seen that these are easier

to use than is the rather irregular surface of the geoid which each is intended to represent. It will be remembered that in ascending order of mathematical complexity (and goodness of fit to a geoid) these are:

- the plane
- the sphere
- the spheroid or ellipsoid of rotation
- the triaxial ellipsoid.

For reasons given in Chapter 1 we do not consider the last of these reference figures, but each of the others are important. We treat with the plane and spherical assumptions in this chapter; with projections of the spheroid in Chapter 16.

The plane assumption

The simplest reference surface is the plane. Position upon it may be described by means of plane rectangular cartesian coordinates, introduced in Chapter 2.

Figure 15.03 illustrates the use of rectangular coordinates in plane surveying. We assume that the origin of the system, O, and the axes, OX and OY have been located, and that we already know the positions of the points $A = (x_a, y_a)$ and $C = (x_c, y_c)$ which are existing control points within the survey. In order to locate another point, B, within the system

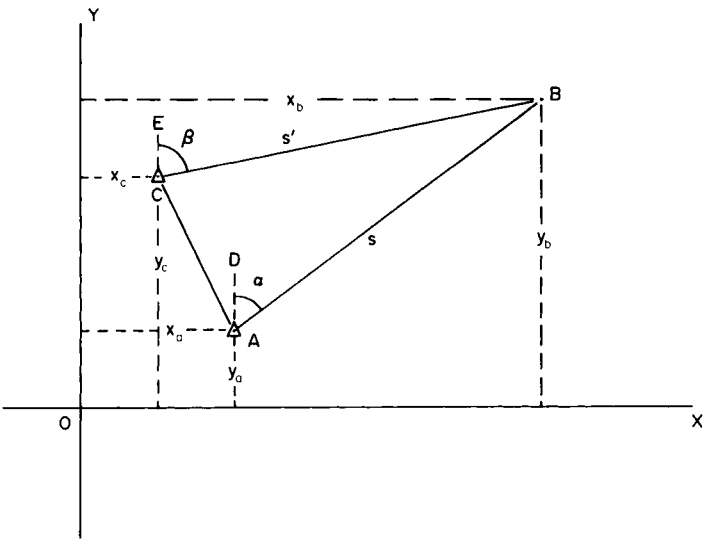


FIG. 15.03 Three survey stations, A, B and C related to one another by plane rectangular coordinates in a system having origin O.

we use the equations

$$x_b = x_a + s \cdot \sin\alpha \tag{15.03}$$

$$y_b = y_a + s \cdot \cos\alpha \tag{15.04}$$

where $AB = s$ and the angle $DAB = \alpha$. Similarly, working from C,

$$x_b = x_c + s' \cdot \sin\beta \tag{15.05}$$

$$y_b = y_c + s' \cdot \cos\beta \tag{15.06}$$

where $CB = s'$ and the angle $ECB = \beta$.

The bearings α and β are derived from combinations of observed angles and known positions. Thus, as A and C have known coordinates, the angle DAC may be determined from any of the equations (2.03), (2.07) or (2.08) on p. 35, and $\alpha = CAB - DAC$.

In a plane survey based upon rectangular coordinates we make certain assumptions about the representation of angles, arcs and distances. We regard the origin of the plane coordinates to be some point near the middle of the survey, although, to avoid having to use negative numbers, the numbering of coordinates is usually referred to a false origin as described in Chapter 2, p. 33. We further assume that we are observing and computing on a plane surface which is tangential to the earth's surface at this point, and that observed spherical angles, such as CAB, are always represented by plane angles of the same size. The observed lines of sight are represented on the plane by straight lines, such as AB, AC and CB, whereas we have seen that these correspond to the arcs of great circles or geodesics upon the curved surfaces of sphere or spheroid. The third assumption is made that plane distances, such as AB, AC and CB have been corrected to represent horizontal distances on the geoid by applying the corrections for slope and, if necessary, height above sea-level. Therefore the plane is a projection in which the principal scale is preserved everywhere and in all directions. The plane assumption has the considerable advantage of simplicity, but since we cannot state categorically that all angles are truly represented, that great circles are straight or that the particular scales behave in a definite fashion without converting equations (15.03)–(15.06) into those for a specific projection, it is desirable to establish practical limits to the size of a survey which can be undertaken using the plane assumption without any significant loss of accuracy arising from this cause. Since this depends upon the accuracy of the measurements which are made, we introduce the somewhat arbitrary practical specifications that:

1. observed directions may contain errors up to 10″;
2. measured distances may be in error by as much as 1 part in 2000.

This corresponds to the standards to be achieved in a small survey job

using a 20″ theodolite and optical methods of distance measurement such as by tachymetry. Where these specifications are satisfied, use of the plane assumption does not materially affect the positions of points provided that the extent of the survey is not more than 10 km. If the accuracy of the observations is lower than those specified, a larger area can be mapped as a plane survey. If the nature of the survey demands a higher order of accuracy, the area for which the plane assumption is valid is correspondingly reduced. However, a great deal depends upon the purpose and ultimate use of the survey. Any survey which is intended solely for the purpose of producing a map or plan needs only to be accurate enough to plot positions within the zero dimension.

For other purposes, such as setting out points on the ground for civil engineering construction work, the precision specified for the initial control has to be carried through to the later stages; therefore the computations must be made with reference to the appropriate projection. The plane assumption is fundamental to much conventional photogrammetry using analogue plotters, because the restitution of photographs is carried out in stereoplotters which operate within the rigid framework provided by three cartesian axes formed by steel bars. We shall consider the special problems involved in this work in Chapter 17.

The spherical assumption

The method of computing the coordinates of a new control point, B, from existing stations such as A and C, is fundamental to surveying practice. If, therefore, the plane assumption is too crude to be justified for a particular purpose and we need to compute the projection coordinates of B, we require a method of doing this from the observed or computed bearings and distances from points whose projection coordinates are already known. In other words it is necessary to obtain equations in x and y which are functions of s and α for the projections used in surveying.

Rectangular spherical coordinates

We must therefore describe another kind of coordinate system for use on the curved surface. These may be called *rectangular spherical coordinates* or *rectangular spheroidal coordinates*, depending upon the reference figure in use. We shall study them here through the mathematically simpler spherical figure, as illustrated in Fig. 15.04. The point O is chosen to serve as the origin of the system, and the ordinate is made to coincide with the meridian through O. We refer to this as the *central meridian*. By analogy with the method of using a plane cartesian reference system, the ordinates of points such as $A = y_a$ and $B = y_b$, correspond respectively to the arc

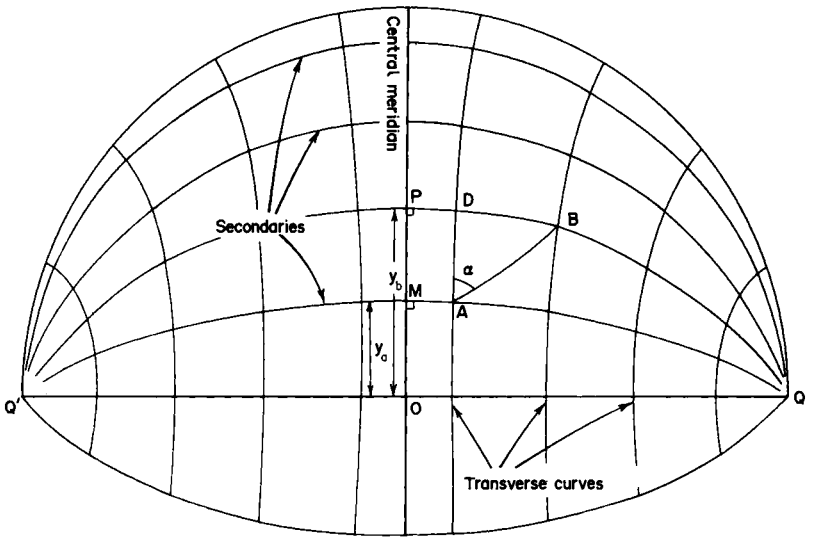


FIG. 15.04 Rectangular spherical coordinates, providing a means of relating points such as A and B, on the curved surface of a sphere to the origin O, by linear distances such as MA and OM.

distances OM and OP along the great circle representing the central meridian. In a plane coordinate system (Fig. 15.03) the abscissae of A and B are the linear distances x_a and x_b. By analogy, therefore, the representation of these distances on the spherical surface are great circle arcs x_a and x_b illustrated in Fig. 15.04. Since we intend to derive *rectangular spherical coordinates*, these two arcs intersect the central meridian at right angles. Therefore in Fig. 15.04, $OMA = OPB = 90°$.

In Chapter 3 it was established, through the definitions of primary great circles, axes and secondary great circles, that meridians must intersect the equator at right angles. Therefore we may argue that the ordinate of the system corresponds to the arc of a primary great circle to which the other great circles such as MA and PB are secondaries. This means that the central meridian must have an axis passing through two poles. It follows that the poles of the system must be the two points Q and Q' which, as shown in Fig. 15.04, are located on the secondary passing through the origin, O, and 90° distant from the central meridian. It follows, therefore, that all great circle arcs which are secondaries to the central meridian converge at Q and Q'.

In plane cartesian coordinates we refer to orientation of any line AB as the bearing, α, defined by the angle DAB in Fig. 15.03. In order to describe the corresponding angle on the spherical surface, we locate the arc AD, in Fig. 15.04, parallel to the central meridian. Thus we employ the convention that $DAB = \alpha$ on the spherical surface. By extension of

the arguments used to describe the properties of geographical coordinates in Chapter 3, DA is the arc of a small circle which is parallel to the primary great circle represented by the central meridian. It follows, therefore, that any small circles thus defined intersect each of the secondaries at right angles. We call the small circles like DA the system of *transverse curves*, and retain the word *secondaries* for the great circle arcs like MA.

The curved surface of the sphere has not been subdivided by the families of secondaries and transverse curves to form a network as shown in Fig. 15.04. Any point on the spherical surface may be related to the origin and axes by its (x, y) coordinates. To define the signs along the axes, we retain the graph convention that $+y$ is towards the North Pole along the central meridian and $+x$ is the direction OQ. We must, however, stress that many writers reverse these directions. This is for reasons given in Chapter 2, pp. 34–35.

The simplest transformation from rectangular spherical coordinates to plane rectangular coordinates is to put in Fig. 15.05:

$$M'A' = MA \text{ and } O'M' = OM$$

which can be expressed algebraically as

$$E_a = x_a \tag{15.07}$$

$$N_a = y_a \tag{15.08}$$

In other words the plane coordinates are made equal to the arc distances along the central meridian and along the secondaries. This is equivalent to the statement that the principal scale is preserved along the central meridian and along the secondaries. It defines a cylindrical projection in

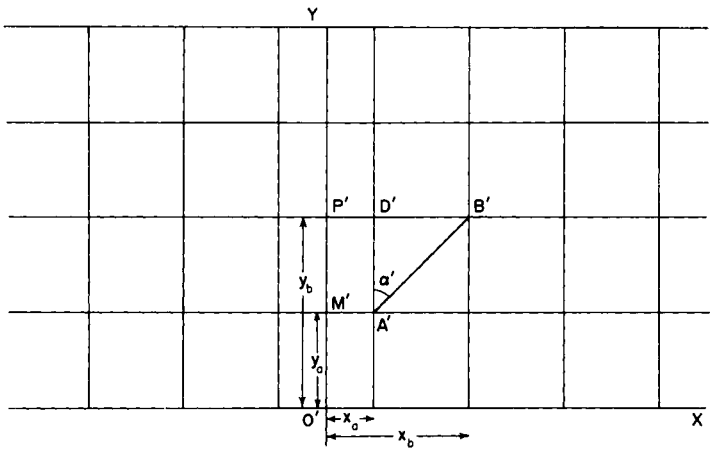


FIG. 15.05 The plane representation of rectangular spherical coordinates by Cassini's projection.

its transverse aspect where the central meridian is a line of zero distortion. Since the principal scale is also preserved everywhere perpendicular to the central meridian, this is the transverse aspect of the *Plate Carrée* or *Cylindrical equidistant*. It is known as the *Cassini–Soldner* projection, or *Cassini's* projection.

The distortions and deformations to be found in Cassini's projection correspond to those of the Plate Carrée (p. 432) but referred to a line of zero distortion along the central meridian rather than the equator. Equation (10.58), p. 212, indicates that the particular scale at the point A' in the direction of the line $A'D'$ is equal to sec z, where z is the angular distance $MA = x_a/R$. Since the principal scale is preserved in the direction $A'M'$, it follows that there is no linear distortion in the east–west direction. Consequently the particular scales at A' vary with direction, and there is angular deformation at A' and every other point which does not lie on the central meridian. In other words, Cassini's projection is not conformal.

Because the linear distortion in the north–south direction increases eastwards and westwards from the central meridian, it follows that Cassini's projection is suitable only for mapping a comparatively narrow zone of longitude. We give expressions for angular and linear distortion in equations (15.12) and (15.16) on pp. 329–330, and Table 15.02 (p. 330) provides numerical values for these. For example, in order to maintain the accuracy specification that errors in direction should not exceed 10″. Cassini's projection can be used only for a zone extending about 80 km either side of the central meridian.

Determination of Cassini coordinates from bearing and distance

There are two ways of finding the projection coordinates of an unknown point B from a known point A, with observed or computed values for $s = AB$ and the bearing $DAB = \alpha$.

In the first method the coordinates (E_a, N_a) are computed directly from the available data; in the second method we apply the corrections to the bearings and distances, and use these corrected values with the ordinary expressions for plane rectangular coordinates (15.03) and (15.04) to find E_b and N_b.

Method I

We proceed from the initial concept that rectangular spherical coordinates are to be represented on the plane by the correct linear distances as specified by equations (15.07) and (15.08). Therefore we treat arcs on the sphere as if they were straight lines on the plane. This can be done if

Legendre's Rule* is applied and the spherical excess of each figure on the spherical surface is calculated. These small angular differences cannot be ignored or discarded. Instead they must be applied as corrections to these equations in the form of second- and third-order terms. Derivation of these terms involves some quite awkward algebra, which we do not attempt to present here. We direct the interested reader to Clark (1944), and similar advanced textbooks on surveying published during the first half of the twentieth century. The final coordinate equations may be written:

$$E_b = E_a + s \cdot \sin\alpha - [s^2 \cdot \cos^2\alpha \cdot E_a/2R^2] - [s^3 \cdot \sin\alpha \cdot \cos^2\alpha/6R^2] \quad (15.09)$$

$$N_b = N_a + s \cdot \cos\alpha + [s \cdot \cos\alpha \cdot E_a^2/2R^2] - [s^3 \cdot \cos\alpha \cdot \sin^2\alpha/6R^2] \quad (15.10)$$

In this pair of equations we note that the first two terms on the right-hand sides correspond to equations (15.05) and (15.06) respectively. In other words, the third and fourth terms in each equation represent corrections to be applied to the plane assumption in order to obtain the Cassini coordinates.

Method II

The alternative way of finding the coordinates of the point B comprises, in effect, introduction of angular distortion and linear deformation to the observed data so that the line AB on the sphere or spheroid is transformed into the line $A'B'$ on Cassini's projection *before the coordinates are calculated.* Figure 15.06 illustrates these corrections, and indicates that we must apply two of them to the bearing α and one correction to the length, s, of the line AB.

The observed line of sight between A and B lies in the plane of the great circle arc passing through these points. Hence the recorded bearing is the great circle bearing DAB, indicated as α in Fig. 15.06. In order to make use of this observation as a plane angle, it is necessary to apply the arc-to-chord conversion which may be needed for quite short lines (roughly speaking all observed lines of length greater than 10 km). By convention, the bearing of the great circle arc, which we have shown in Fig. 15.06 as α, is also denoted by t. The bearing of the rhumb-line, or chord, which we call α_0, is sometimes denoted by T. Hence this correction is often known to surveyors as the *t–T correction.* The angle $\alpha - \alpha_0 = t - T$ is frequently also denoted by δ, and since on the spheroidal surface the amount of correction differs at each end of the line, we must further distinguish between δ_{AB}, which is the arc-to-chord conversion to be

*Legendre's Rule may be stated as follows: 'If one-third of the spherical excess of a spherical triangle is deducted from each angle, the triangle may be solved in terms of the linear lengths of the sides by the ordinary rules of plane trigonometry.'

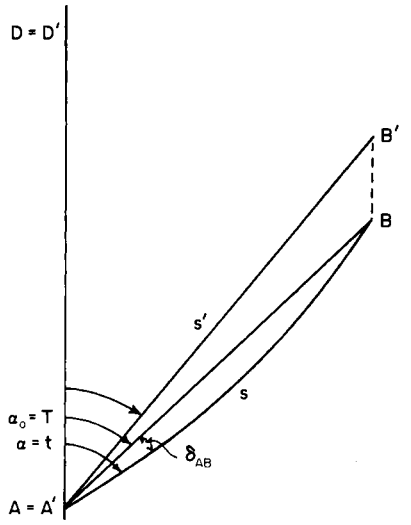


FIG. 15.06 The relationships between the angles and sides measured on the ground and their representation on Cassini's projection. This diagram attempts to compare measurements made on the sphere with their plane representation as follows. The arc AB is measured on the ground and is found to be of length s with bearing $\alpha = t$ referred to grid north through A. Application of the arc-to-chord conversion, δ_{AB} (which the diagram indicates is also equal to $t-T$ or $\alpha-\alpha_0$, gives the chord bearing, α_0 of the point B referred to grid north through A. The point B' is the position of B on Cassini's projection. This has distance $A'B' = s'$ from the point A' and the line $A'B'$ bears β measured from grid north through A'. Hence Method II, or point-to-point working, requires calculation of the bearing, β, and the distance s', in order to find the Cassini coordinates for B'.

applied to the bearing measured at A, and δ_{BA}, which is the corresponding correction to be applied to the bearing measured at B. The correction to be applied has the form

$$\delta_{AB} = (t-T)'' = [(N_b-N_a)(E_b+2E_a)]/6R^2 . \sin 1'' \qquad (15.11)$$

where δ_{AB} is expressed in seconds of arc.

The direction in which the correction is applied, in other words the sign of δ_{AB}, depends upon the orientation of the line AB with respect to the central meridian, as shown in Fig. 15.07. The magnitude of the correction depends upon the length of the line and the bearing.

The second correction which must be applied to the bearing is that needed to convert the plane angle DAB into the plane angle $D'A'B'$ on the projection. Since Cassini's projection is not conformal, we expect there to be angular deformation which is equivalent to a rotation of the chord AB towards the central meridian. Denoting $D'A'B' = \beta$ and since

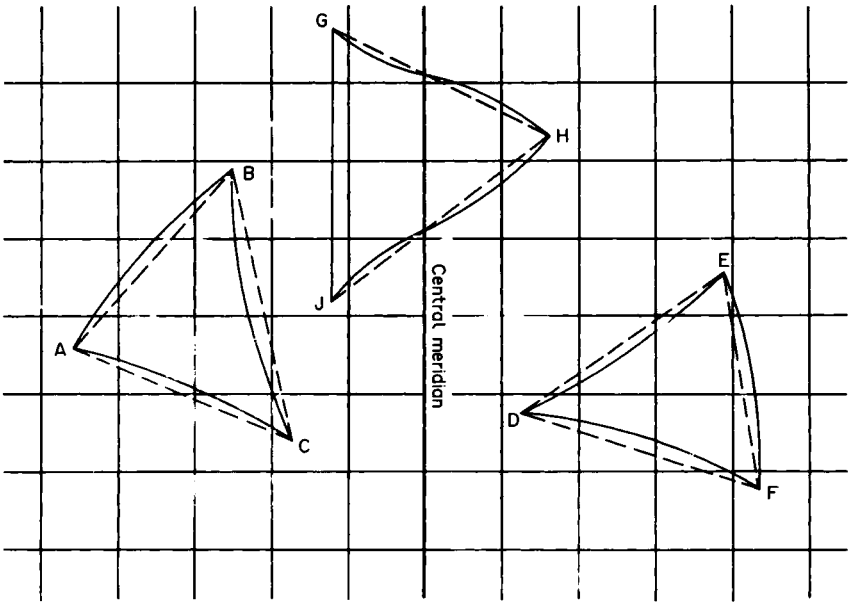


FIG. 15.07 The nature of the arc-to-chord conversion for six different lines having different positions and orientation with respect to the central meridian of Cassini's (and the Transverse Mercator) projection. Note that the arc and chord only coincide along the line *GJ* which is parallel to the central meridian, or, in other words, a line having a grid bearing of 360° or 180°.

$DAB = \alpha_0 = T = \alpha - \delta_{AB}$, this correction has the form

$$(\beta - \alpha_0)'' = [(\sin\alpha_0 . \cos\alpha_0)/6R^2 . \sin 1'']E_\mu^2 \qquad (15.12)$$

where

$$E_\mu^2 = (E_a^2 + E_a E_b + E_b^2) \qquad (15.13)$$

Thus to find the required bearing β from the observed angle α, the corrections are applied in the following order:

$$\alpha + \delta_{AB} = \alpha_0 \qquad (15.14)$$

$$\alpha_0 + (\beta - \alpha_0) = \beta \qquad (15.15)$$

Equation (15.12) is expressed in seconds of arc but both (15.11) and (15.12) can be expressed in radians by dropping the sin 1″ term. The two angular corrections can be combined in a single equation but we prefer to keep them separate for in the conformal Transverse Mercator projection equation (15.12) is equal to zero and only (15.11) remains.

The linear correction to be applied to an infinitely short line $AB = ds$

to convert this into $A'B' = ds'$ may be found from the equation

$$ds'/ds = 1+(\cos^2\alpha_0/6\,R^2).\,E_\mu^2 \qquad (15.16)$$

Finally, therefore, the coordinates of B' may be expressed by the two equations

$$E_b = E_a+s'.\sin\beta \qquad (15.17)$$

$$N_b = N_a+s'.\cos\beta \qquad (15.18)$$

which are, of course, in the same form as equations (15.05) and (15.06).

We may now use equations (15.16) and (15.20) to determine the maximum distortion to be expected in the use of Cassini's projection. We solve these equations for the directions in which α_0 has the maximum effect. Thus in (15.16) the greatest deformation in bearing occurs where $\alpha_0 = 45°$, 135°, 225°, and 315°. In (15.20) the greatest linear deformation occurs where $\alpha_0 = 0°$ and 180°, corresponding to the conclusion on p. 326 that the particular scale $\mu_1 = \sec z$ is directed along the projection of the transverse curves parallel to the central meridian. Numerical values for the maximum distortions are given in Table 15.02.

The unsuitability of Cassini's projection for mapping a whole country having the size of Britain in a single unit may be gauged by assuming Britain to have been mapped on this projection using the same origin and central meridian as the National Grid. Then in longitude 8°40′W, or 6°40′ from the central meridian, which is in the vicinity of St Kilda, the maximum linear distortion would be 1/530 and the maximum deformation in bearing 3′14″. From equations (15.12) and (15.13) we can see that the deformation in bearing $(\beta-\alpha_0)$ is independent of length of line, for the right-hand side of the equation only contains arguments in Eastings. Thus in the Outer Hebrides a line of length 10 m would be deflected through about 3′ of arc, as is a line of length 10 km or 100 km. Even for the most rough-and-ready kind of survey this amount of angular deformation would be intolerable. The remedy adopted in the earliest days of the Ordnance Survey was to use a different origin and central

TABLE 15.02 *Maximum distortions in distance and bearing for Cassini's projection*

	Distance in km from central meridian					
	50	100	150	200	250	300
Maximum distortion in distance	1/32 500	1/8100	1/3600	1/2000	1/1300	1/900
Maximum distortion in bearing	3″	13″	29″	51″	1′19″	1′54″

meridian for each county. This is described on pp. 334–335 and illustrated in Fig. 15.09.

Geographical coordinates on Cassini's projection

Although it has been suggested that most survey computations may be made without reference to geographical coordinates, we cannot complete the description of this projection without any reference to the equations which are used to relate geographical coordinates to grid coordinates.

Cassini's projection is here derived for a sphere of radius R. Treatment of projections of the spheroid are left to the consideration of the Transverse Mercator in Chapter 16. If anybody still needs the coordinate expressions for Cassini's projection of the spheroid, they will find it in the first edition of this book.

Figure 15.08 illustrates the relationship between the Cassini coordinates and the geographical coordinates of the point A' on the projection corresponding to A on the spherical surface. In Fig. 15.08(b), O' is the origin of the (E, N) system of grid coordinates and $N'O'$ is the axis representing the central meridian (λ_0). In Fig. 15.08(a), NA is the meridian λ_a through A. The angle ONA therefore corresponds to the difference in longitude $\lambda = \lambda_a - \lambda_0$. DA is parallel to the central meridian and therefore indicates the direction of grid north. The angle DAN is the convergence at A. The parallel of latitude φ_a through A meets the central meridian at F. We denote the latitude of the point M by φ' and we shall call it the *foot-point latitude*. Then

$$\varphi_a = m/R \tag{15.19}$$

where m is the meridional arc distance from the origin O to F. For the sphere this corresponds to the distance s_m in (3.10) if the origin, O, is located at the equator, or $s_{m''}$ in equation (3.12) if the origin of the projection is in some other latitude. Moreover

$$\varphi' = N_a/R \tag{15.20}$$

and

$$z = E_a/R \tag{15.21}$$

It is required to express E_a and N_a in terms of φ_a and λ.

From the right-angled triangle NMA

$$\sin z = \cos \varphi_a \,.\, \sin \lambda \tag{15.22}$$

$$\tan \varphi' = \tan \varphi_a \,.\, \sec \lambda \tag{15.23}$$

Substituting (15.20) and (15.21) in the left-hand side of (15.22) would provide us with equations in E_a and N_a but not in a form suitable for

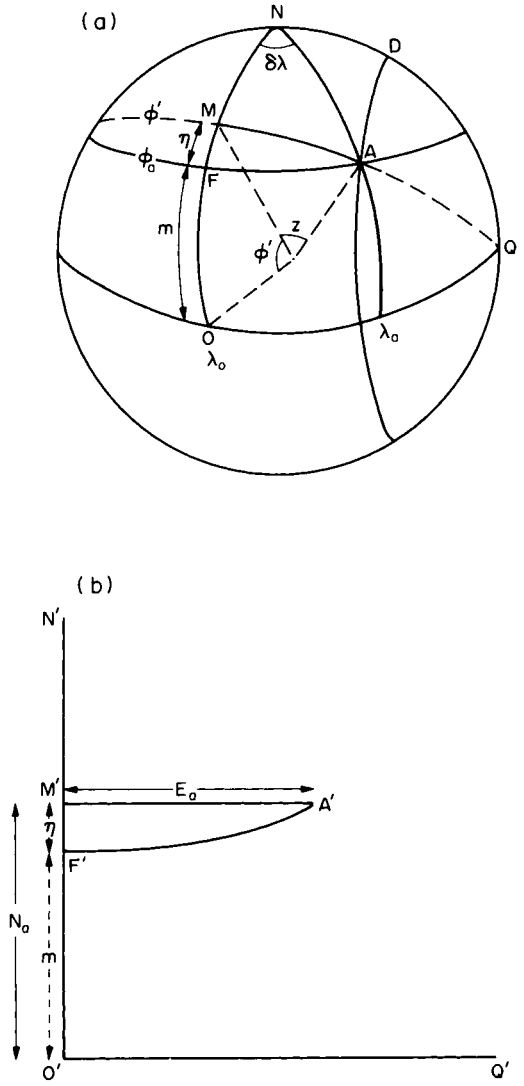


FIG. 15.08 The relationship between the Cassini coordinates and geographical coordinates of a point.

easy computing. Therefore it is desirable to transform equations (15.22) and (15.23) by expanding certain of the terms in series using well-known mathematical techniques, which were first introduced in Chapter 4. Applying those for $\sin \lambda$ and $\sin^{-1} z$, namely

$$\sin \lambda = \lambda - \lambda^3/6 + \lambda^5/120 - \ldots \tag{15.24}$$

and

$$\sin^{-1} z = z + z^3/6 + 3z^5/40 \ldots \tag{15.25}$$

equation (15.22) may now be written in the form

$$z = \cos\varphi_a[\lambda - \lambda^3/6 + \lambda^5/120 - \ldots] + (1/6).\cos^3\varphi_a[\lambda - \lambda^3/6 + \lambda^5/120 - \ldots]^3 + (3/40).\cos^5\varphi_a[\lambda - \lambda^3/6 + \lambda^5/120 - \ldots]^5 \tag{15.26}$$

It has already been shown that $z = E_a/R$. Therefore equation (15.26) may be written as an expanded expression to find E_a:

$$E_a = R.\lambda.\cos\varphi_a + (R/6)(-\tan^2\varphi_a)\lambda^3.\cos^3\varphi_a + (R/120)(-8\tan^2\varphi_a + \tan^4\varphi_a)\lambda^5.\cos^5\varphi_a \tag{15.27}$$

A certain amount of additional algebra is still needed to derive the Northings equation. In (15.23) we put

$$\tan\varphi' - \tan\varphi_a = \tan\varphi_a(\sec\lambda - 1) \tag{15.28}$$

Moreover we find use for the curious circular argument that

$$\varphi_a = \tan^{-1}(\tan\varphi_a) \tag{15.29}$$

Expansion by Taylor's Theorem gives

$$\varphi' = \varphi_a + (\tan\varphi' - \tan\varphi_a).\cos^2\varphi_a - (\tan\varphi' - \tan\varphi_a)^2\cos^4\varphi_a.\tan\varphi' \tag{15.30}$$

We now substitute in (15.28) the series corresponding to sec λ. The result is substituted for $(\tan\varphi' - \tan\varphi_a)$ in (15.30) and since $\varphi_a = m/R$, we finally obtain the equation

$$N_a = m + \tfrac{1}{2}R.\tan\varphi_a.\lambda^2\cos^2\varphi_a + (R/24)\tan\varphi_a(5 - \tan^2\varphi_a)\lambda^4.\cos^4\varphi_a \tag{15.31}$$

From Fig. 15.08(b), $N_a = O'M' = O'F' + F'M'$. Since $O'F' = m$, the linear distance $F'M'$ may be expressed by the second and third terms of the right-hand side of (15.31) or

$$\eta = \tfrac{1}{2}R.\tan\varphi_a.\lambda^2\cos^2\varphi_a + (R/24)\tan\varphi_a(5 - \tan^2\varphi_a).\lambda^4.\cos^4\varphi_a \ldots \tag{15.32}$$

This quantity was often referred to as the *ordinate of curvature* in the older literature on the subject, but the use of this term seems to be no longer fashionable. The use of the lower-case Greek eta (η) to represent this variable was also common. However, this term is also used to simply the algebra of the Transverse Mercator projection of the spheroid, and in Chapter 16 we shall encounter its use in an entirely different context, thereby leading the unwary astray.

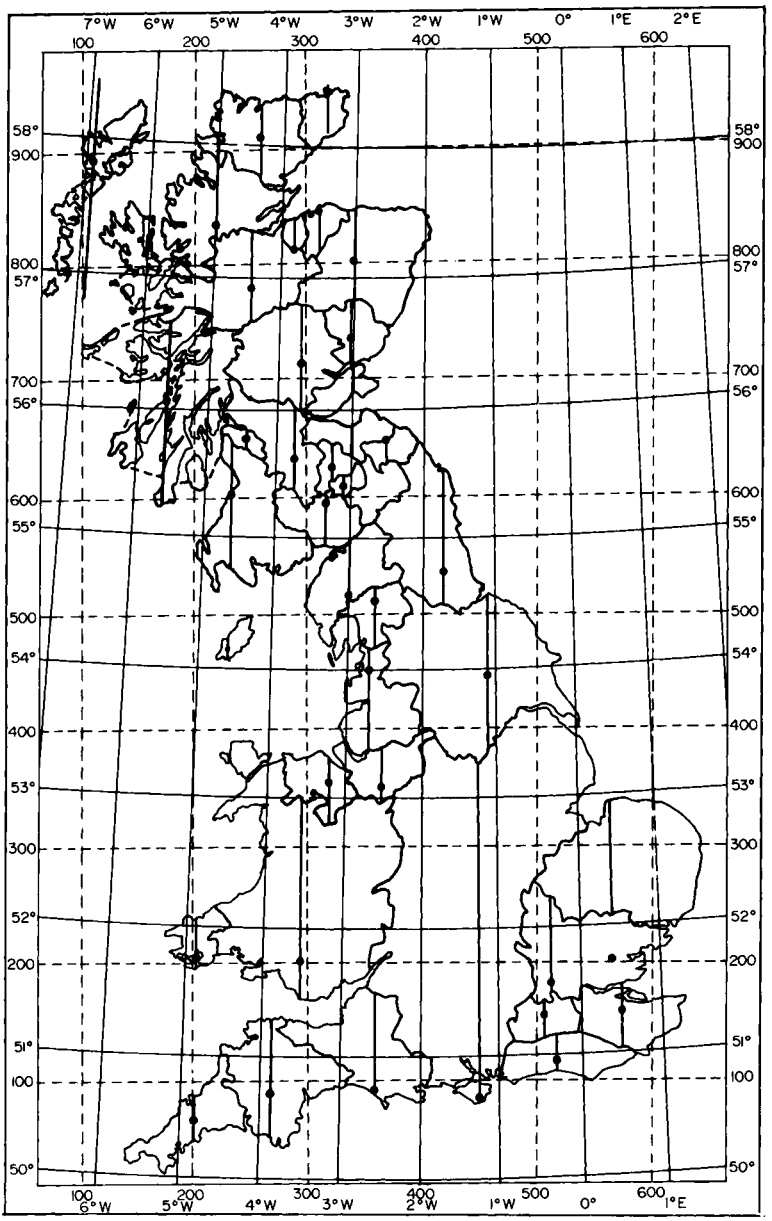


FIG. 15.09 The origins, central meridians and boundaries of the parts of Britain mapped separately on different versions of Cassini's projection before 1938. The dots indicate the origin of each projection system and are primary Ordnance Survey triangulation stations. The central meridian of each system is indicated by a thick line. The map also show the National Grid of the unified projection system which is also illustrated in Fig. 2.02.

Cassini's projection and the Ordnance Survey

At the time when the first national surveys were created in Britain and France opinion about the objects of their work differed from current views. Contemporary surveying policy was to carry out small and independent surveys of individual counties, communes, parishes and towns, rather than the creation of an integrated national survey. Winterbotham (1934) quotes from documents as late as the middle of the nineteenth century which maintained that the map of each parish and each county should be complete in itself.

Consequently the early work of the Ordnance Survey emphasised the subdivision of Britain into counties to the extent that all maps and plans of scale 1/10 560 (Six Inches to One Mile) or larger were created as separate *County Series*, with additional sets of larger-scale plans for certain towns. Therefore each country or group of counties was mapped on a separate Cassini projection, each having its own origin and central meridian. The essential simplicity of Cassini's projection, expressed by equations (15.07) and (15.08), favoured the use of it for these small areas. Even after some regrouping of the projections in use, which took place towards the end of the nineteenth century, the maps of Great Britain were still based upon 39 different Cassini projections, as shown in Fig. 15.09. Hotine (1947) commented upon the result as follows

> Now the Cassini projection – which still enjoys a quite unmerited popularity in some parts of the world – does not even begin to fill the bill. Even for the most rough and ready purposes it cannot be extended very far without correction, and the corrections to observations are difficult without in effect transforming to another projection. But if the projection is limited to too small an area then we encounter too often the difficulty of a junction – of transforming from one system to another. . . . For instance, when the Cassini projection was adopted in England – I may say at a time when no other system was readily available – it was considered good enough to cover an English county and separate projections were laid down for each county. The county was the administrative unit whose boundaries were thought most unlikely even to change. Unfortunately they have changed and in addition such entities as the town of Sheffield have quite irreverently sprawled across them, with the result that frequent transfers of large-scale surveys from one system to another have been necessary, in some cases with most unfortunate results. In fact we have got into such a mess that it has necessarily been decided to scrap the lot and put the whole country on a single projection system.

In Britain the change to a conformal projection coincided with the retriangulation, which was carried out during the 1930s and completed after World War II, so that the Transverse Mercator coordinates of trigonometric control points are based upon a new and entirely independent series of observations (OS, 1967). Most of the features of the National Grid, which is the plane grid upon which the Ordnance Survey Transverse Mercator projection is constructed, have already been mentioned in Chapter 2.

CHAPTER 16

The Transverse Mercator projection

> It will go so far to suggest that if the problem had ever been considered solely on the merit of practical application, we should not now be using the Transverse Mercator projection at all. What is the use of incurring complexity in order to achieve a rigorous meridian scale condition when we immediately throw it away by applying an overall scale factor?
>
> M. Hotine, *Empire Survey Review*, 1946

Introduction

Because of the great importance of the Transverse Mercator projection it is desirable to consider it in detail and distinguish between the different versions of it. As elsewhere in this book, excepting Chapter 10, it has been the author's intention to avoid much of the algebra involved in the derivation of a projection. This applies especially to the study of the different versions of the Transverse Mercator projections of the spheroid. Plenty of other people have already done this, generally to the exclusion of other information concerning the history of its use and practical ways of organising the calculation of the various equations. The most complete study of the Transverse Mercator projection is still to be found in the work of the Bulgarian geodesist V. K. Khristov (whose name is often written Hristow, which is the German transliteration from the Cyrillic form). His monumental book, *Gauss-Krüger Coordinates on the Ellipsoid of Rotation*, was originally published in Germany during World War II (Hristow, 1943) and the Russian-language version appeared more than a decade later (Khristov, 1957). The book was translated into English by the US Army Map Service, but this was evidently never published. In addition, Khristov published 40 other papers on the subject in the *Zeitschrift für Vermessungswesen* between 1934 and 1944, some of which are listed as items 1249 through 1272 in Snyder and Steward (1988). As a sample of the English-language contributions the reader is referred to the work of Lee (1945), Hotine (1946–7), Redfearn (1948) and Jackson (1978, 1980). Moreover there are still new ideas to put forward, for example in the papers by Williams (1982), Agajelu (1987), Day (1990) and Bowring (1989, 1990a, 1990b).

There are various ways in which the Transverse Mercator formulae for the spheroid may be derived.

- The first is to derive an additional term for the Eastings equation to convert the Cassini coordinates given in equation (15.09) into Transverse Mercator coordinates, a procedure which Jackson (1978) has dismissed as: 'that sloppy, unmathematical statement that, by adding a term or two to the Easting formula of the Cassini projection one can produce a transverse orthomorphic construction. This is a strange way to define a projection'. Notwithstanding this criticism, the author still considers a description of this method to be the logical development from rectangular spherical coordinates into conformal mapping. Moreover, it demonstrates one of the simplest methods of point-to-point working and is therefore practically useful.
- The second is a direct representation of the ellipsoid on the plane, generally in an attempt to preserve lengths along the central meridian. The theory of the projection was developed by Gauss between 1820 and 1830, who used it for the original control survey of Hanover at that time. It was further studied by Dr L. Krüger in 1912, who presented equations in a form suitable for logarithmic solution. Consequently it is usually known as the *Gauss–Krüger projection.*
- The traditional approach to the Gauss–Krüger projection has now been modified considerably by Bowring (1989), who has replaced some of the terms in the rather complicated equations for mapping the spheroid by simpler expressions referring to the sphere. This modification is essentially based upon new ways of looking at the algebraic manipulation of the Gauss–Krüger equations and is not to be confused with the following version.
- This is known as *double-projection*, or sometimes the *Gauss–Schreiber projection.* The transformation is carried out in two stages; first conformal representation of the ellipsoid upon a sphere of appropriate radius; secondly conformal mapping of the sphere to a plane. There are several ways in which this may be done, too.

The Transverse Mercator projection of the sphere

We shall proceed as the derivation from the Cassini projection of the sphere as described in the previous chapter, pp. 325–327 and equations (15.07)–(15.10).

The Transverse Mercator is a conformal projection in which the line of zero distortion is a meridian. Therefore the principal scale is preserved along this meridian and, referring back to Fig. 15.05, p. 325:

$$OM = y_a = O'M' = N_a = \varphi' . R \qquad (16.01)$$

as in the Northings equation for Cassini's projection (15.08). It follows, moreover, that the particular scale along the transverse curves is still $\mu_1 = \sec z$, where, as before

$$z = E_a/R \tag{16.02}$$

The requirement is for a conformal projection of the sphere. Therefore the particular scales at the point A' must be equal in all directions. If we substitute z for φ throughout, we may use the same arguments as those presented in Chapter 10 to derive the normal aspect of Mercator's projection of the sphere, and expressed by equations (10.53)–(10.73). It follows that the particular scale is equal to sec z and the Eastings equation for the point A' is

$$E_a = R \,.\, \ln \tan (\pi/4 + z/2) \tag{16.03}$$

Hence equations (16.01) and (16.03) define the Transverse Mercator

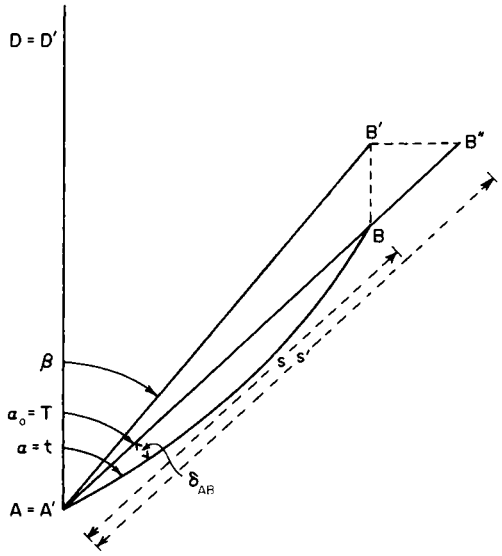


FIG. 16.01 The relationship between the angles and sides measured on the ground and their representation on the Transverse Mercator projection. Compare this with Fig. 15.06, p. 328, for the definitions of the bearings α, α_0 and β on Cassini's projection. The point B is the ground point which has been distance s and bearing α from the ground point A. The point B' is, as in Fig. 15.06, the position of B on Cassini's projection. The point B'' represents the position of B on the Transverse Mercator projection. Note that the angular correction $\alpha_0 - \beta$ does not apply to the Transverse Mercator coordinates, which corresponds to the attempt to make this a conformal projection. The only corrections to be applied are the arc-to-chord or $(t-T)$ correction and the determination of the distance $A'B' = s'$ on the projection.

coordinates of a point in terms of z and φ', though this is not a particularly useful way of doing it.

The relationship between the Cassini and Transverse Mercator coordinates of a point B may be illustrated by means of Fig. 16.01. Since the Transverse Mercator projection is conformal, it follows that

$$\beta - \alpha_0 = 0$$

or

$$\beta = \alpha_0 \tag{16.04}$$

In other words, the bearing $D'A'B'$ on the plane corresponds to the rhumb-line bearing DAB. Consequently the line AB is projected as $A'B''$ whereas it was $A'B'$ on the Cassini projection. Moreover the Northing of B'' is the same as that for B', corresponding to the initial conditions presented above. Therefore equation (15.10) for Cassini's projection remains unchanged on the Transverse Mercator projection of the sphere. The required modification to the Eastings equation is that representing the distance $B'B''$. This can be shown, for example in Clark (1973) to be

$$B'B'' = E_B'^3/6R^2 + E_B'^5/24R^4 \tag{16.05}$$

where E_B' are the Cassini Eastings coordinates computed from equation (15.09). Therefore the coordinates of a point B determined by the equivalent to Method I are, for the Transverse Mercator projection of the sphere

Plane element | *Cassini element*

$$E_B = E_A + s \cdot \sin\alpha - s^2\cos^2\alpha \cdot E_A/2R^2 + s^3 \sin\alpha \cdot \cos^2\alpha/6R^2$$

| *Transverse Mercator element*

$$+ E_B'^3/6R^2 + E_B'^5/24R^4 \tag{16.06}$$

$$N_B = N_A + s \cdot \cos\alpha + s \cdot \cos\alpha \cdot E_A^2/2R^2 - s^3 \cdot \sin^2\alpha \cdot \cos\alpha/6R^2 \tag{16.07}$$

The terms on the right-hand sides of equations (16.06) have been labelled to show how they comprise a *plane element*, a *Cassini element* and a *Transverse Mercator element*. It must be emphasised that this relationship is valid only for a projection of the sphere.

In conventional surveying applications equations (16.06) and (16.07) are less useful. The preferred technique corresponds to Method II described in Chapter 15. This is much simplified for Transverse Mercator coordinates because the term $(\beta - \alpha_0)$ of equation (15.12) does not have to be calculated. Consequently there remain only the arc-to-chord conversion and a correction for the distance, s. For the Transverse Mercator projection, the expression for the $(t-T)$ correction, in (15.11) is still valid. The alteration to the equation for correcting linear distance (15.16) results from the fact that the particular scales at any point in a conformal

projection are constant in all directions. It follows that $\alpha_0 = 0$, hence

$$\cos^2 \alpha_0 = 1$$

and the corresponding equation is

$$ds'/ds = 1 + E_\mu^2/6R^2 \tag{16.08}$$

where

$$E_\mu = (E_a^2 + E_a E_b + E_b^2) \tag{16.09}$$

for a line AB. For an infinitely short distance from the point, the quantity ds'/ds is the scale-factor, which we have seen in Chapter 15 is same as the particular scale at a point. It follows that the errors quoted in Table 15.02 as the *maximum* scale errors on Cassini's projection represent the linear error *in any direction* on the Transverse Mercator. Table 15.02 further indicates that the use of the projection should be confined to a narrow zone either side of the line of zero distortion which, in a transverse aspect projection, is a narrow zone of longitude either side of the central meridian. However, we may reduce excessive linear distortions towards the edge of a zone by introducing the simple form of modification already described in Chapter 11. Just as it is possible to modify the normal aspect conical or cylindrical projection by introducing the concept that the principal scale is preserved along two standard parallels, so in transverse cylindrical projections the principal scale may be preserved along a pair of transverse curves which are equidistant from the central meridian. In a conformal projection this modification can be introduced by using a scale factor $k_0 < 1{\cdot}0$. Then (16.08) may be written

$$ds'/ds = k_0[1 + (E_\mu^2/6R^2)] \tag{16.10}$$

For practical use in systems where the zone extends about 3° either side of the central meridian, a suitable scale factor is $k_0 = 0{\cdot}9996$, which corresponds to a reduction in scale of 2499/2500 along the central meridian. This scale factor has been used for both the Ordnance Survey version of the Transverse Mercator and the UTM system. Reduction of the principal scale by 0·9996 has the effect of creating two lines of zero distortion at a distance of about 180 km from the central meridian.

A variety of other values for k_0 have been introduced for particular purposes. For example in the different versions of the Transverse Mercator projection used for the State Coordinate Systems in the USA each has been specially tailored to provide the best projection within the boundaries of a particular State; to ensure that the scale errors of the projection do not exceed 1/10 000. It should be remembered that differences of this size can be detected by EDM, so that this does not represent an impossibly high standard of accuracy which is beyond the measuring capabilities of the highway or civil engineer.

From the general formulae developed in Chapter 10 to describe all cylindrical projections, it seems logical to apply a corresponding scale factor to Cassini's projection. This can be done, although the idea does not seem to have occurred to nineteenth-century surveyors. It was left for Young (1920) to suggest that the modification could be applied, and he claimed that he was first to demonstrate this. By then, however, the importance of Cassini's projection was already waning fast. In any case the task of recomputing the whole of a regional or national control network was, in those days, such a formidable undertaking that the cost of conversion would barely justify the benefits. Consequently this method of improving Cassini's projection was never used.

The geometry of the Transverse Mercator projection of the spheroid

We have considered the geometry of the projections used for surveying purposes and topographical mapping purely in terms of the projection of

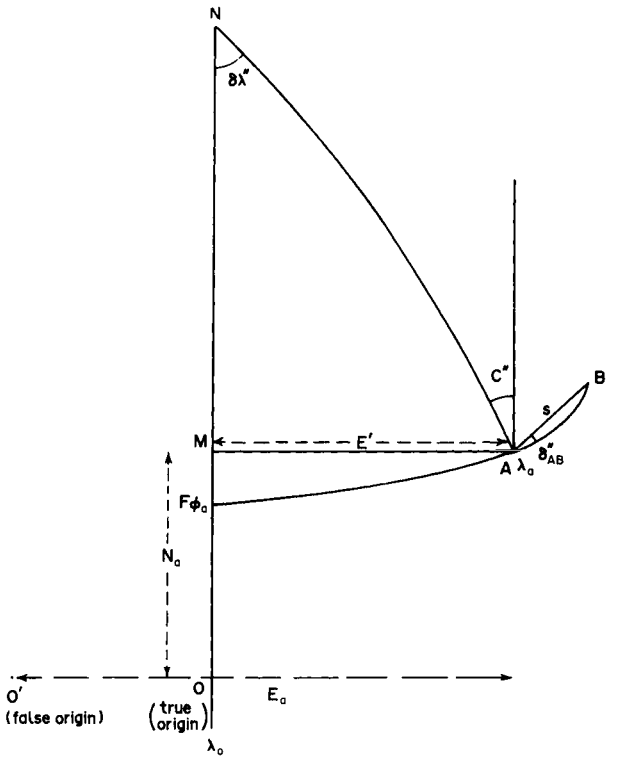


FIG. 16.02 The geometry of the Transverse Mercator projection illustrating the principal components which may require to be calculated.

a spherical figure; because this has the advantage, already emphasised in Chapter 5, that the solutions by spherical trigonometry are exact and represented by closed expressions. The expanded versions of the Cassini and Transverse Mercator equations are used for the greater ease of computation using projection tables, logarithms and mechanical calculators. We have seen that the elliptic functions describing the ellipsoid of rotation cannot form closed expressions and the mapping equations have to be represented as series, usually in ascending powers of λ.

In order to execute all the computations which may be required in surveying and cartography, we refer to the list of solutions which was given on pp. 70–73.

Direct representation of the ellipsoid upon the plane: The Gauss–Krüger projection

Consider a point having geographical coordinates (φ, λ) where the longitude λ is reckoned from the central meridian. Then N is the plane coordinate distance measured along the central meridian from the origin of the projection, which may be where the central meridian intersects the equator or, in special cases, some other parallel of latitude reckoned to pass through the origin. As in Chapter 15, E is the distance of this point from the central meridian, reckoned along the geodesic which is perpendicular to this meridian. The basic formulae for Northings and Eastings are derived as series in powers of λ expressed in radians.

We observe that:

- When λ is zero, the formula for E must be zero, and the formula for N must simply be N = m, where m is the meridional arc distance along measured from the equator (or, as in some national projections, from the true origin). See Chapter 4, equations (4.16) and (4.20) for methods of determining m for the spheroid.
- Since the projection is geometrically symmetrical with respect to the central meridian, the formula for N must be unchanged when $+\lambda$ is changed to $-\lambda$, and the formula for E must change sign without change of numerical value. Hence

$$N = m + P.\lambda^2 + Q\lambda^4 + R\lambda^6 + \ldots \tag{16.11}$$

$$E = A.\lambda + B\lambda^3 + C\lambda^5 + \ldots \tag{16.12}$$

with only even powers for λ in N and odd powers in E. In practice λ is a small angle, usually not exceeding 3°, or about 0·05 radians. It follows that the higher powers of λ, such as λ^3, λ^4, λ^5, λ^6, form a succession of extremely small numbers. The coefficients A, B ... P, Q ... are functions of the latitude φ, they have the dimension of length, and they incorporate those parameters which define the

spheroid to which the survey system is referred, namely the radii of curvature, ρ and v and either the first or second eccentricity e^2 or e'^2.

From the elementary account of the differential geometry of the sphere given in Chapter 5, pp. 94–99, it is easy to transform equation (5.12) from the spherical version to the spheroidal simply by introducing the radii of curvature for the spheroid. Thus the length of the infinitely short arc AB on the surface of the spheroid is given by

$$(ds)^2 = \rho^2(d\varphi)^2 + v^2\cos^2\varphi(d\lambda)^2 \qquad (16.13)$$

differing only from (5.12) by the use of the radii of curvature v and ρ in place of the spherical radius R.

The corresponding distance on the projection is ds′, or

$$(ds')^2 = (dN)^2 + (dE)^2 \qquad (16.14)$$

and it follows that the particular scale at A' in the direction $A'B'$ is ds′/ds.

On substituting the differential formulae for dN and dE we get

$$(ds'/ds)^2 = \{[(\partial N/\partial\varphi)^2 + (\partial E/\partial\varphi)^2](d\varphi)^2 + 2[\partial N/\partial\varphi \,.\, \partial N/\partial\lambda\varphi \,.\, \partial E/\partial\lambda] + \partial\varphi \,.\, \partial E/\partial\lambda(d\varphi)(d\lambda) + [(\partial N/\partial\lambda)^2 + (\partial E/\partial\lambda)^2](d\lambda)^2\}/\{\rho^2(d\varphi)^2 + v^2 \,.\, \cos^2\varphi \,.\, (d\lambda)^2\} \qquad (16.15)$$

which is the version of equation (5.31) for the spheroid.

It is at this stage that two special conditions are introduced; first that the projection should be conformal; secondly that the central meridian shall be represented by its true length throughout. The special property of conformality requires that the particular scales should be independent of the direction of an arbitrary line AB as specified in Chapter 6, pp. 106–107, and already considered on p. 338 of this chapter. The second condition may be simply written

$$N = m \qquad (16.16)$$

where m is the meridional arc distance from the equator, or from some arbitrary defined origin.

Having established these conditions, one of the commonest ways of proceeding is to convert from geodetic into isometric latitude, q. This was originally described on pp. 67 and 216, equation (10.83). The reader is reminded that

$$q = \ln\tan(\pi/4 + \varphi/2)[(1 - e \,.\, \sin\varphi)/(1 + e \,.\, \sin\varphi)]^{e/2} \qquad (10.83)$$

At this stage we also introduce the use of complex variables into the work. This represents another big step in the mathematical competence expected of the reader, and we must take this without adequately explaining the methods used. Those who already known about complex variables

will find the application here elementary enough. Those who know little of the subject are referred, in particular, to the study of conformal projections by Thomas (1952), where the theory of complex numbers is introduced in some detail with special reference to the kind of coordinate transformation needed here. A recent paper by Bowring (1990b) relates specifically to application of complex numbers to the Transverse Mercator projection.

Using complex variables, the general solution of the equation for a conformal projection is

$$x+iy = f(q+i\lambda) \tag{16.17}$$

where i is the complex number ($i^2 = -1$).

It should be noted that in equation (16.17) and those which follow the (x, y) coordinate axes are reversed so that the x-axis corresponds to the central meridian. The reason for this was explained on p. 34.

The first of these conditions demands that, when $y = 0$, then x shall be a function of q alone, and therefore $\lambda = 0$; that is, the x-axis or initial meridian must be selected as the origin of longitude.

If we select the central meridian as origin of longitude, then the first condition is satisfied, for when $\lambda = 0$,

$$x+iy = f(q) \tag{16.18}$$

which requires that $y = 0$.

The second condition demands that, when $y = 0$, then $x = m$, where m is the arc of the meridian from the equator (or any other chosen origin) to the point (x, 0). But when $y = 0$, then $x = f(q)$, and hence

$$f(q) = m \tag{16.19}$$

is the required condition. The rest of the derivation is, as Hotine once described it, 'a matter of brute force and algebra'.

The function on the right-hand side of (16.17) is now expanded using Taylor's Theorem in a series of ascending powers of $i\lambda$, to produce

$$x+iy = f(q)+i\lambda(df(q)/dq)+[(i\lambda)^2/2!][d^2f(q)/dq^2] + [(i\lambda)^3/3!][d^3f(q)/dq^3]+ \ldots \tag{16.20}$$

whence, by equating the real and imaginery parts on either side of this equation, and remembering that $f(q) = m$, the coordinates are given by

$$x = m-\tfrac{1}{2}\lambda^2(d^2m/dq^2)+\lambda^4/24(d^4m/dq^4)- \ldots \tag{16.21}$$

$$y = \lambda(dm/dq)-\lambda^3/6(d^3m/dq^3)+\lambda^5/120(d^5m/dq^5)- \ldots \tag{16.22}$$

However, these are expressed in isometric latitude whereas we require geodetic latitude. The successive derivatives of m are readily obtained with the help of (10.83) and differentiation with respect to φ, leading to

the following equations.

$$x = m + \tfrac{1}{2}\lambda^2 . \nu . \sin\varphi . \cos\varphi + 1/24\lambda^4 \nu \sin\varphi . \cos^3\varphi (5 - \tan^2\varphi + 9e'^2 . \cos^2\varphi) + \ldots \quad (16.23)$$

$$y = \lambda . \nu . \cos\varphi + 1/6\lambda^3 . \nu . \cos^3\varphi (1 - \tan^2\varphi + e'^2\cos^2\varphi) + 1/120\lambda^5 . \nu . \cos^5\varphi (5 - 18\tan^2\varphi + \tan^4\varphi + 14e'^2\cos^2\varphi^2 - 58e'^2\sin^2\varphi) \ldots \quad (16.24)$$

where the second eccentricity, $e'^2 = (a^2 - b^2)/b^2 = e^2/(1 - e^2)$, and the higher powers of e'^2 have been discarded. These series are rapidly convergent, and in practice it is only the first two or three terms of each expression that are needed.

Although equations (16.23) and (16.24) represent usable expressions for determining E and N respectively, they do not represent the only solution. There are a number of small algebraic adjustments which may be applied to make the equations look shorter, neater and more elegant, or even make them easier to compute. In addition to those used to describe the spheroid, which were given in equations (4.03)–(4.07), on p. 65, a most commonly used substitution is that

$$\eta^2 = e'^2\cos^2\varphi = \nu/\rho \quad (16.25)$$

The reader should be aware that this use of η has nothing to do with the geometry of the meridional ordinate of curvature as employed in Chapter 15. From (16.25) the various equations for the Transverse Mercator projection of the spheroid may be expressed in various combinations of ν/ρ, η and t (where $t = \tan\varphi$). Also one of the most elementary of trigonometric substitutions is that

$$\tan\theta = \sin\theta/\cos\theta,$$

so that t may also be expressed in terms of $\sin\varphi$ and $\cos\varphi$. The version used for the expanded series in the various equations depends upon little more than personal preference. Redfearn (1948) provided all the equations in two versions; the 'η and t' and the 'ν/ρ and t'. His equations, upon the following are based, included the terms to λ^6 and λ^7 and likewise, E^6 and E^8.

We do not interrupt the narrative of this chapter by listing any of these equations in detail. They are to be found in Appendix III, pp. 443–449, where they may be compared with one of the nested arrangements of the equations which are much more economical to use in computer programs or for solution by desk calculator.

Bowring's solution

We have seen that the derivation of the geographicals–grid solution is in the form of (16.11) and (16.12), this being the logical way of arranging the solutions when these have to be done using projection tables and each term is multiplied by some power of longitude, λ. However, we do not use tables any more in these surveying solutions so that the constraint on how the equations are organised need no longer apply. A recent study of the Transverse Mercator projection by Bowring (1989) has involved reworking the algebraic methods used by Redfearn (1948) with the aim of simplifying the solutions by replacing the non-elliptic terms with a spherical solution. Apart from the obvious merits of the method in writing suitable programs to determine Transverse Mercator parameters, it provides an interesting bridge between the pure Gauss–Krüger solution and the double-projection methods. Here we comment briefly upon Bowring's solution of equation (III.2), corresponding to (16.24), which is to find E from φ and λ.

Bowring's paper provides alternative solutions for all the standard equations in Appendix III and through the determination of six spherical angles, θ_1 through θ_6. We indicate here only the derivation of θ_1.

Starting from equation (III.2), and after two pages of algebraic manipulation, Bowring offers the following equation for the Eastings.

$$\mathrm{E} = k_0 \,.\, \nu[\ln \tan(\pi/4 + \theta_1/2) + 1/60 \,.\, \mathrm{e}'^2 \,.\, \lambda^3 \,.\, \cos^5 \varphi(10 - 29\lambda^2 + 36\lambda^2 \,.\, \cos^2 \varphi)] \quad (16.26)$$

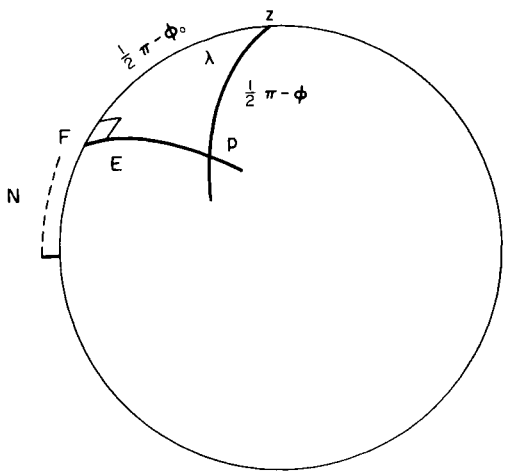


FIG. 16.03 The geometry of Bowring's solution, showing the spherical triangle from which the angles θ may be determined.

TABLE 16.01 *Transverse Mercator projection; Bowring's solution of the geographicals-grid calculation. Comparison of the terms in the Eastings coordinate calculated for middle latitudes and the International Spheroid. The scale factor $k_0 = 1{\cdot}0$ and the tabulated values for Eastings terms are expressed in metres*

Latitude φ	Longitude λ	Full solution		Eastings	
		E (m)	N (m)	Spherical term (m)	Spheroidal term (m)
30	1	96 492·784	3 320 582·756	96 492·7654	0·0186
	2	193 000·615	3 321 846·380	193 000·4654	0·1492
	3	289 537·688	3 323 953·931	289 537·1846	0·5033
	4	386 119·960	3 326 907·683	386 118·7673	1·1924
40	1	85 397·885	4 430 084·018	85 397·8751	0·0101
	2	170 801·220	4 431 521·566	170 801·1397	0·0876
	3	256 213·195	4 433 918·717	256 212·9226	0·2722
	4	341 640·162	4 437 277·328	341 639·5176	0·6442
50	1	71 699·433	5 541 438·040	71 699·4285	0·0042
	2	143 393·141	5 542 876·238	143 393·1073	0·0336
	3	215 079·822	5 545 273·961	215 079·7091	0·1131
	4	286 756·698	5 548 632·293	286 756·4304	0·2673
60	1	55 801·460	6 654 650·131	55 801·4586	0·0020
	2	111 594·375	6 655 915·401	111 594·3653	0·0096
	3	167 370·373	6 658 024·398	167 370·3409	0·0322
	4	223 120·487	6 660 977·447	223 120·4106	0·0759

We see that the first ($\log_e$ tan) term in parentheses is, of course, the right-hand side of (16.03) for a sphere of radius $k_0 . \nu$ and Bowring's equation to find θ_1

$$\theta_1 = \sin^{-1}(\sin\lambda . \cos\varphi) \tag{16.27}$$

is the same as equation (15.22) which we derived for Cassini's projection of the sphere.

Thus the Eastings equation (16.26) comprises a spherical term together with a correction to convert this to the spheroid. Bowring claims that within the range of longitude $-4° < \lambda < 4°$, this equation agrees with the full Redfearn solution of (III.2) to within 0·5 mm on the ground. Table 16.01 provides some results for middle latitudes, and demonstrates that within the normal longitude range of $\pm 3°$ for Transverse Mercator zones, the spheroidal correction barely exceeds 0·5 m and is often only a few millimetres.

The Transverse Mercator by double-projection

This version appears to date from Schreiber, who first used it for the control surveys established by the Preussische Landesaufnahme in the

mid-nineteenth century as an alternative to the direct form of projection originally described by Gauss. The Institut Géographique Nationale used a form of double-projection known as the *Gauss–Laborde projection*, for mapping French African colonies (AOF and AEF) between 1947 and 1951. However, this was soon replaced by the UTM. It follows that the Gauss–Schreiber versions of the Transverse Mercator have never been as important as the Gauss–Krüger projection, and until recently might have been considered an historical curiosity. However, the recent study of the Transverse Mercator projection, by Williams (1982) and Agajelu (1987) has demonstrated that the method has some merit, and it may be conveniently modified for use with microcomputers and even pocket calculators. The disadvantages of the Gauss–Schreiber projection have been partly overcome by the introduction to small corrections which would have been inconvenient to present in tabular form in the days when the projection coordinates had to be determined by logarithms or clumsy mechanical calculations.

As we have seen, this conformal representation is made in two stages; first of the ellipsoid to the sphere and secondly of the sphere to the plane. An early attempt to obtain the Gauss–Krüger projection without using complex variables led McCaw (1940) to reinvent the method of double-projection, but Lee (1945) found fatal errors in McCaw's reasoning so that the projection was not even conformal. As a result of this detailed criticism of McCaw's work, Lee's paper is a most valuable introduction to the concept of double-projection. Nevertheless, it is important to realise that the difference between the incorrect and correct versions is extremely small and limited to a few of the higher-order terms. Table 16.01 has shown that the differences between the Eastings term for the sphere and spheroid, as determined by Bowring, are usually very small; in the example of McCaw's double-projection and Lee's equations for the Gauss–Krüger projection they are even smaller.

Conformal projection from the ellipsoid to the sphere

In order to map the curved surface of the spheroid upon that of a sphere, the first requirement is to establish the relationship between the geographical coordinates (φ, λ), which we here describe as the *geodetic coordinates* of a point on the spheroid into the corresponding angles (Φ, Λ) on the sphere. There are a series of different transformations which can be made to correspond to the different mappings of the spheroid. We have already referred to these in Chapter 4, p. 67. Since we need to make a conformal map of the spheroidal surface, the transformation (φ, λ) $\rightarrow$ (Φ, Λ) is conformal.

The second requirement is to establish a suitable radius R for the sphere to correspond to the radii of curvature ρ, ν of the spheroid. Several

different solutions have been proposed; that owing to Gauss, the so-called *Biernacki–Rapp* solution; that proposed by Hotine and the most recent by Williams. We cannot devote the space here to treat with them in detail, but refer the reader to the relevant literature.

We start from the three equations of differential geometry which have already been given elsewhere. These relate to the arc ds, which is expressed in terms of latitude and longitude:

On the sphere

$$ds_S^2 = R^2 d\varphi^2 + R^2 \cos^2\varphi \,.\, d\lambda^2 \tag{5.12}$$

On the spheroid

$$ds_E^2 = \rho^2 d\varphi^2 + v^2 \cos^2\varphi \,.\, d\lambda^2 \tag{16.13}$$

On the plane

$$ds'^2 = dx^2 + dy^2 \tag{5.18}$$

or, as in equation (16.14)

$$ds'^2 = dE^2 + dN^2 \tag{16.28}$$

We now introduce the new variable

$$dq = [\rho/(v \,.\, \cos\varphi)]d\varphi \tag{16.29}$$

which, after substituting terms in e and φ from the expressions (4.08) and (4.09), eventually becomes

$$dq = (1/\cos\varphi) \,.\, d\varphi + \tfrac{1}{2}e\{[-e(1+e\sin\varphi)\cos\varphi - e(1-e \,.\, \sin\varphi)\cos\varphi]/ [(1-e \,.\, \sin\varphi)/(1+e \,.\, \sin\varphi)](1+e \,.\, \sin\varphi)^2\} \tag{16.30}$$

Integrating this expression

$$q = \ln\tan(\pi/4 + \varphi/2) + \tfrac{1}{2}e\ln[(1-e \,.\, \sin\varphi)/(1+e \,.\, \sin\varphi)] + C \tag{16.31}$$

If we set the condition that for $\varphi = 0$, $q = 0$, the integration constant, $C = 0$ and we may write

$$ds^2 = v^2 \,.\, \sin^2\varphi(dq^2 + d\lambda^2) \tag{16.32}$$

The system (q, λ) are a system of isometric coordinates and q is the isometric latitude.

On the sphere

$$q = \ln\tan(\pi/4 + \varphi/2)$$

whereas for the spheroid (10.83) applies.

Finally, the radius of the sphere is taken to be the quantity

$$R = [\rho_0 \,.\, v_0]^{1/2} \tag{16.33}$$

which is the radius of the *Gauss mean sphere*. For a derivation of (16.33) see Richardus and Adler (1972).

The Biernacki–Rapp solution

This name is used by Agajelu (1987) to describe the solution published by Biernacki (1965) as taught by Rapp (1975) in the Department of Geodetic Science at Ohio State University.

If we define as (Φ, Λ) the coordinates on the sphere which correspond to (φ, λ) on the spheroid then

$$\Phi = P_0 + P_2\lambda^2 + P_4\lambda^4 + \ldots \tag{16.34}$$

$$\Lambda = P_1\lambda + P_3\lambda^3 + P_5\lambda^5 + \ldots \tag{16.35}$$

in which the terms in P are coefficients to be derived from expansions in n and φ. Obviously these terms have geometrical significance. For example P_0 corresponds to the meridional arc length on the conformal sphere. The individual equations used to determine P_i are listed here in Appendix III (III.57)–(III.62) on p. 449. It remains to determine the radius of the conformal sphere. In order to impose the condition that $\varphi = \Phi = 90°$, and that the scale along the central meridian is unchanged, we must put

$$m = P_0 . R \tag{16.36}$$

where, as before, m is the meridional arc distance on the spheroid. Using an expansion in n, we may write for the meridional arc distance for the quadrant Q from $\varphi = 0$ through $\varphi = 90$

$$Q = m_{90} = [a/(1+n)](1 + n^2/4 + n^4/64 + \ldots) . \pi/2 \tag{16.37}$$

and

$$R = [a/(1+n)](1 + n^2/4 + n^4/64 + \ldots) \tag{16.38}$$

Williams' solution

This was published in Williams (1982). On the parallel where the two surfaces are tangential, the curvature of the meridional arc on the spheroid and that on the sphere are obviously equal, and there is a belt of latitude within which meridian arcs on the sphere and the spheroid are virtually indistinguishable. Therefore we may write

$$R(\varphi' - \varphi'_0) = m \approx m_0$$

or

$$R\Delta\varphi' = \Delta m \tag{16.39}$$

At the heart of Williams' method is the more exact solution for (16.39),

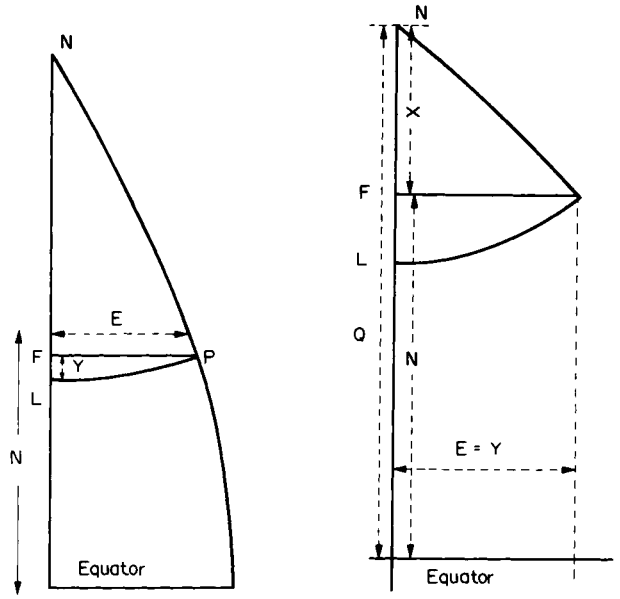


FIG. 16.04 Williams' solutions for the Transverse Mercator projection.

namely

$$R\Delta\varphi' = \Delta m - u(\Delta m)^4 - v(\Delta m)^5 - w(\Delta m)^6 - \ldots \qquad (16.40)$$

where

$$u = e'^2 . \sin\varphi_0 . \cos\varphi_0 / 6\rho_0 v_0^2 \qquad (16.41)$$

$$v = [e'^2 \cos^2\varphi_0 / 30\rho_0^2 v_0^2][1 - (6e'^2 \sin^2\varphi_0)/v_0] \qquad (16.42)$$

$$w = [e'^2 \sin\varphi_0 \cos\varphi_0 / 180\rho_0^3 v_0^2][-2 + 3\tan^2\varphi_0 + (2e'^2\rho_0(3 - 13\cos^2\varphi_0))]/[v_0 + (39e'^2\rho_0^2 \sin^2\varphi_0 . \cos^2\varphi_0)/v_0^2] \qquad (16.43)$$

Williams' method for obtaining the geographicals-to-grid equations is illustrated in Fig. 16.04. He also developed the transformation expressions for use in high latitudes (where the above solution is weak). This is based upon placing the origin of the projection at the geographical pole and working equator-wards along the central meridian rather than the customary practice of working pole-wards from an origin at the equator.

Hotine's solution using the aposphere

At about the same time that Lee investigated the errors in McCaw (1940), Hotine was working on his monumental work, listed as Hotine (1946–7), together with the investigation of suitable conformal projections for

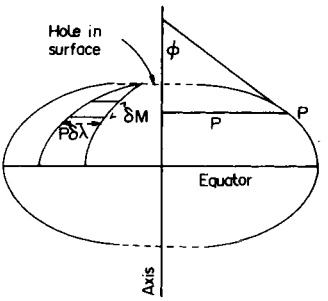


FIG. 16.05 The aposphere. (Source: Bomford, 1962.)

countries such as the Federated Malay States and the British colonies of British North Borneo and Sarawak, as there were in those days. The result was the *Rectified Skew Orthomorphic* projection, later described by Brazier (1951) and which is now called the *Hotine Oblique Mercator* or HOM projection by the Americans.

As the intermediate surface between the spheroid and plane, Hotine made use of a deformable body known as an *aposphere*. The importance of this has been described in Hotine (1946–7), Bomford (1980) and Thompson (1973, 1975).

Projection of the sphere to the plane

Having determined conformal latitude and longitude on a sphere of suitable radius R, it remains to project the point with coordinates (Φ, Λ) conformally to the plane. The closed formulae of Thomas (1952) are appropriate

$$E = R \tan^{-1}(\tan \Phi . \sec \Lambda) \tag{16.44}$$

$$N = (R/2) \ln[(1 + \cos \Phi . \sin \Lambda)/(1 - \cos \Phi . \sin \Lambda)] \tag{16.45}$$

The history of use of the Transverse Mercator

The following section describes the use of the principal versions of the Transverse Mercator projection in chronological order of their introduction.

The Gauss–Krüger projection in Germany

Notwithstanding the early use of the projection for control surveys by Gauss and Schreiber, by the end of the nineteenth century most of the large-scale map cover of Germany was mapped on one or other variants

TABLE 16.02 *Gauss–Krüger projection: Local scale-factors in middle latitudes, International Spheroid, $k_0 = 1{\cdot}0$. Values determined from the nested equation (III.19), p. 445, by Meade (1987)*

Latitude φ	Longitude λ	Scale-factor k	Latitude φ	Longitude λ	Scale-factor k
30	1	1·0001142	50	1	1·0000631
	2	1·0004594		2	1·0002524
	3	1·0010340		3	1·0005663
	4	1·0018390		4	1·0010096
40	1	1·0000897	60	1	1·0000381
	2	1·0008078		2	1·0001525
	3	1·0008078		3	1·0003431
	4	1·0014365		4	1·0006098

of the polyhedric projection. The first modern use of the Gauss–Krüger projection as a national system was its introduction to Germany in 1923 to replace these. In the modern application the Transverse Mercator projection used in Germany comprised a succession of zones 3° wide (i.e. $1\frac{1}{2}$° either side of the central meridian) and with a scale-factor of unity. Although the scale-factor at the edges of a 3° zone is appreciably smaller than the corresponding values at the edge of a 6° zone, the narrowness of the 3° zone creates the difficulty that much work is located inconveniently near the boundary between two zones; therefore additional transformations are required each time a line of sight crosses from one zone to another. Consequently there is conflict of interests between the needs of surveys for municipal, cadastral and civil engineering purposes, for which small projection distortions are desirable and the needs of military topographical work, where the inconvenience caused by frequent zone changes is more important than the small loss in accuracy towards the edges of a 6° zone. In Germany the 3° cadastral zones were modified for military use by extending them to the 6° zones during World War II.

Soviet unified reference system

In 1928 the third geodetic conference of the State Planning Commission of the USSR decided to adopt a national projection system comprising the Gauss–Krüger projection in zones of width 6° with specified central meridians and zone boundaries for a world-wide system.

The specifications for the system, which has been retained to the present, is a uniform cover of the world in zones which correspond to the 6° longitudinal units of the *International Map of the World*, and which were also the same zones chosen for the Universal Transverse Mercator (UTM) system 20 years later. Because there has been no form of private land ownership in the USSR since 1917, there are obviously no cadastral

requirements to be considered. Municipal surveys and large civil engineering undertakings are referred to local Transverse Mercator systems. The projection is the direct Gauss–Krüger projection of the spheroid to the plane. The origin for each zone is at the intersection of the equator with the central meridian. This point is assigned the arbitrary Eastings value of 500 000 m to avoid, as in other systems, having to use negative Eastings. There is no alteration in the scale on the central meridian so that this corresponds to the principal scale = 1·0 as in the German version of the projection, and this distinguishes them from the UTM where a scale factor of 0·9996 is used. Fister (1980) has made a detailed comparison between the Gauss–Krüger and UTM versions to demonstrate that, apart from the different Figures of the Earth and scale-factors employed, there is no other significant difference between them. The only major change which has been introduced to the Soviet Unified Reference System since its initiation has been to change the spheroid. In the 1920s and 1930s the Bessel ellipsoid was still used for surveys of the USSR, but the change was made in 1942 to the Krasovsky spheroid. During the post-war period the system was introduced to all the Eastern bloc countries and it was even applied to those parts of Antarctica where the Russians were active from 1956 onwards.

The Ordnance Survey Transverse Mercator projection

A Departmental Committee on the Ordnance Survey was appointed in 1935, under the chairmanship of Viscount Davidson, to investigate future policy. The so-called Davidson Committee presented its final report in 1938, and its recommendations were eventually accepted by HM Government in 1945 as a guide for the post-war policy of the department.

One of the most important of the recommendations contained in this report was the replacement of the patchwork of County Series of medium- and large-scale maps, based upon the many Cassini projections illustrated in Fig. 15.09, by a single national projection system tied to the unique reference system of the National Grid. All future Ordnance Survey control was to be located on a Gauss–Krüger version of the Transverse Mercator projection using the origin, axes and numbering convention of the National Grid described in Chapter 2.

The principal arguments concerning this choice of projection which were given in the Davidson Report are also summarised in Ordnance Survey (1950, 1967), and Seymour (1980). In fact the introduction of the projection for the computation of the primary retriangulation antedated the Davidson Committee by several months, and it was employed for this purpose many years in advance of all other applications. The maximum difference in longitude in Britain (excluding Ireland which has its own

TABLE 16.03 *Scale-factor and linear distortion encountered in the use of the Transverse Mercator projection of Great Britain used by the Ordnance Survey*

	Distance from central meridian in km					
	0	50	100	150	200	250
Nation Grid Easting	400 km	350/450 km	300/500 km	250/550 km	200/600 km	150/650 km
Local Scale factor	0·9996	0·99963	0·99972	0·99988	1·00009	1·00037
Linear Distortion	1/2500	1/2700	1/3570	1/8330	1/11110	1/2700

version of the Transverse Mercator projection) is more than 10°, from 1°43′E near Great Yarmouth to 8°34′W on St Kilda. The true origin of the National Grid is, we have seen in Chapter 2, the point in latitude 49″N, longitude 2°W. Hence the use of the Ordnance Survey projection extends from about 3°43′ eastwards and 6°34′ westwards from the central meridian, which is more than twice the maximum difference in longitude possible in the use of the UTM and SUR systems. It follows that there are some fairly large linear distortions present in the Outer Hebrides. Without any modification this would be of the order of 1/531, but the projection is modified by the introduction of a scale-factor of 0·9996 on the central meridian, which reduces the distortion to 1 part in 675. This does not appear to be a matter of much concern in Britain, though in most other countries this would be regarded as justification for introducing a second Transverse Mercator zone. We have already seen that the arc–chord correction can also be quite large at such a distance from the central meridian. Of course these are extreme values. The great majority of surveying in Britain is done on the mainland where distances from the central meridian are much less. However, continental shelf surveys might extend even further west than St Kilda. We do not wish to emphasise the importance of this island itself, but to direct attention to the potential economic value of the surrounding seas. Use of the OS projection as the basis for mapping the seabed here might be considered to be stretching the limit of this system rather further than was ever envisaged, and it is likely that either the Irish version of the projection or the UTM are to be preferred as far west as this.

Both of the Irish national surveys (the Ordnance Survey of Ireland and the Ordnance Survey of Northern Ireland) use a separate projection. This has precisely the same properties as that described, being based upon the same Airy 1830 Figure of the Earth and employing a scale-factor of 0·9996 on the central meridian. The only difference is the origin, which

is situated in latitude 53°30′N, longitude 8°00′W.* This origin is located near the centre of the Republic of Ireland. It is therefore necessary to locate the false origin in latitude 51°13′N, longitude 10°32′W, or 200 km west of the central meridian and 250 km south of the latitude of the origin. Consequently Eastings in the Irish version of the Transverse Mercator have the constant 200 000·0 m and the Northings of the Irish version have the constant 250 000·0 m added to them. In fact, UTM zone 29 has its central meridian in longitude 9°W, which serves very nicely for mapping the whole of Ireland. Tables 16.03 indicates the scale distortions which may be experienced in the use of the Ordnance Survey version of the Transverse Mercator projection.

The African dilemma

During the interwar years there was much discussion, but very little action, concerning the production of maps for British colonial territories in Africa. There was virtually no disagreement about the type of projection to be used, for the suitability of the Transverse Mercator projection was accepted very early. The arguments centred on two aspects of the mapping process: the units of measurement to be used, which we do not consider here, and the width of the Transverse Mercator zones to be adopted. Since the surveying requirements for land registration in South Africa have always been notoriously strict, and since the cadastral requirement for farm surveys was the principal impetus to map production in that country, the Union of South Africa took the lead both in the introduction of the Gauss–Krüger projection and in the use of zones of width 2°, where the scale factor never exceeded 1·000 13 or about 1/8000 (for a difference in longitude of 1° in latitude 25°S).

In most of the African colonies the major stumbling block was that the cadastral surveyors insisted upon the need for narrow zones, but the military topographical surveyors wished to have fewer grid zones within a given territory and therefore preferred the wider, 6°, zones. With hindsight it seems remarkable that the possibility of having two systems in use, one for cadastral mapping and one for topographical mapping, was never seriously considered. The history of the controversy has been described by McGrath (1976), who has commented.

> It is an unavoidable conclusion that these unresolved differences continued to be seen and accepted as the irreconcilable views of the cadastral surveyor and topographic surveyor.
>
> ... Was it necessary to force uniformity in cadastral (grid) systems? If this had not been attempted, it is entirely possible that a uniform projection and grid zone for

*We observe that St Kilda lies close to the central meridian of the Irish projection, but is referred to that for England, Wales and Scotland.

topographic mapping might have been accepted throughout Southern, Central and East Africa almost twenty years before it came about. Moreover the often intense mistrust of the (military) topographic surveyor by the (colonial) cadastral surveyor could have been mitigated.

The subsequent history is revealed in the published discussions of the first two postwar Commonwealth Survey Officers Conferences held in 1947 (CSO, 1947) and 1951 (CSO, 1951). With the creation of the Directorate of Colonial Surveys in 1946 a concerted attempt was made to reconcile the two requirements (still trying to adhere to the idea of a single projection system) and a compromise had been achieved by adopting 5° zones. However, as Hotine in the 1951 discussion commented

> The nearest we could get at the last (1947) Conference was acceptance of the 5° Transverse Mercator Belts for the Commonwealth territories South, Central and East Africa. We have always realised, however, that this stood no chance of acceptance by our French, Belgian and Portuguese neighbours. Yet many of the problems we have to face in Africa are international in character and do need consistent series of maps to assist their solution.

Colonel Baumann, who was at that time the Director of Trigonometrical Survey in the Union of South Africa, gave vent to his feelings in no uncertain fashion (CSO, 1951):

> At the last Conference in 1947, South Africa was accused of being the 'nigger in the woodpile'. I was told that, had we not been so stubborn, a 5° Gauss Grid would have already covered most of Africa...
>
> The conference ultimately decided to adopt 5° grids for Africa for Topographical Survey purposes, provided the majority of states accepted this proposal, and provided further that there was some guarantee of permanency.
>
> I accepted this resolution in good faith and upon my return to South Africa immediately set in train the work of transforming twenty thousand fixed points from the 2° to 5° Gauss belts. The transforming of these twenty thousand fixed points from one co-ordinating system to another would appear to be child's play to our friends in America who, we have heard, think nothing of recomputing the co-ordinating of eight hundred thousand points in an incredibly short space of time, but to South Africa with its limited resources this work was a considerable undertaking. I am now being asked to throw away all this work overboard and start again. It was quite by chance that I heard of this proposed change....
>
> I am quite prepared to play ball with the Americans on UTM provided I am assured that it will be accepted by the majority of states in Africa south of the Sahara, and provided, further, that I have a firm undertaking that this is the last change which the South African Government will be called upon to make in the sphere of Grids for Topographical Survey purposes.

The Universal Transverse Mercator (UTM) system

The UTM comprises the following features

- The projection is the Gauss–Krüger version of the Transverse Mercator intended to provide world coverage between the latitudes 84°N and 80°S.
- The unit of measure is the International Metre.

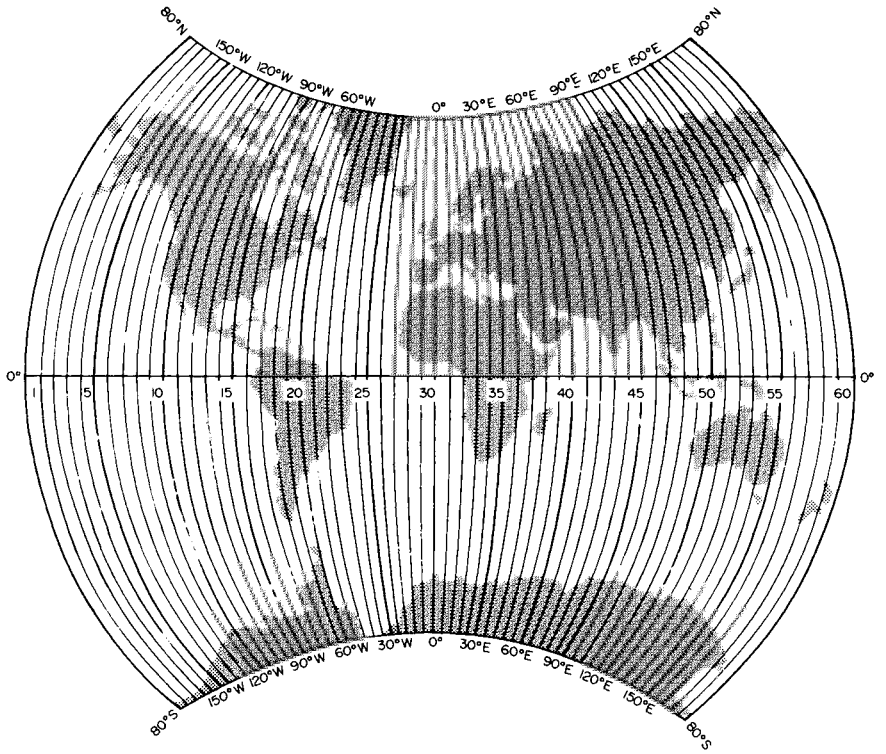


FIG. 16.06 The method of zone numbering according to the UTM system. Each zone is 6° longitude in width and extends from 80°S to 80°N.

- Each zone is 6° of longitude in width; the first zone has its western edge on the meridian 180° and the zones proceed eastwards to zone 60 which has its eastern edge at 180° longitude. The central meridian of each zone is therefore 177°W in Zone 1, 171°W in Zone 2 and 168°W in Zone 3. The succession of zones is illustrated in Fig. 16.06.
- The origin of each zone is the point on the equator where it is intersected by the central meridian of the zone.
- Each zone extends as far pole-wards as 84°N and 80°S. Initially these limits were set at 80°N and 80°S.
- The Eastings of the origin of each zone is assigned the value of 500 000 m.
- The UTM has a special convention for recording Northing coordinates in the southern hemisphere. For the southern hemisphere the equator is assigned a northing value of 1 000 000 m, but this is zero in the northern hemisphere. Figure 16.06 illustrates the coordinate numbering conventions.

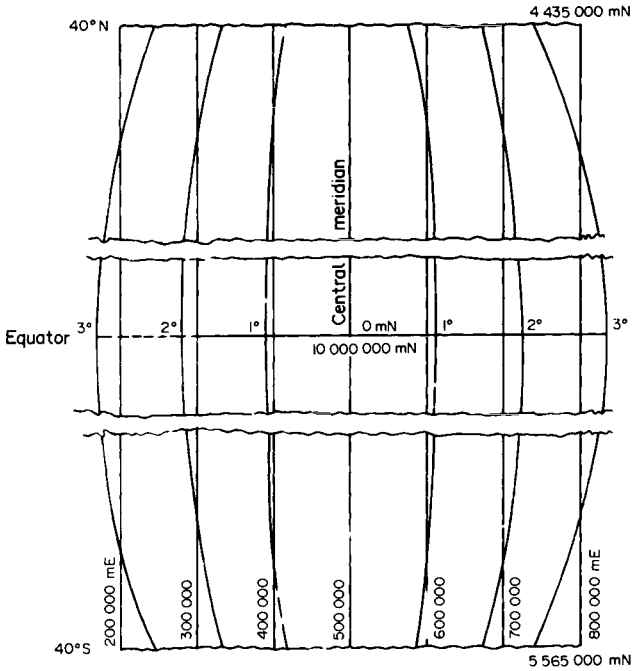


FIG. 16.07 The relationship of grid to graticule in any single UTM zone. This diagram is arbitrarily truncated near latitudes 40°N and 40°S.

- The scale-factor on the central meridian is 0·9996.
- The UTM employs five different figures, for specified areas. The greatest extent of cover is that provided by the International Spheroid; Clarke 1866 is used for Africa; Clarke 1880 for North America; Everest and Bessel for smaller parts of Southern and South-east Asia.

The UTM had been devised in the late 1940s by the US Army, who then began to introduce it as the base for all military mapping used by NATO and other western military alliances. The event which prompted the irritable response by Colonel Baumann at the 1951 Commonwealth Survey Officers Conference was the discussion of the paper by Hough (1955), from the US Army Map Service, who had the unenviable task of presenting the proposals for the introduction of the Universal Transverse Mercator System. He had a very rough ride in justifying the suitability of UTM to some of those attending the 1951 Conference of Commonwealth Survey Officers as the record of the discussion on this paper testifies.

Baumann was not the only person to object to the way the system had been introduced. For example, in the same discussion (CSO, 1951), Brigadier J. C. T. Willis, who at that time was Director of Military Survey in Britain, stated:

> The impact of UTM on military mapping has been the subject of prolonged discussion with the United States and ourselves for several years. The object of signing some such agreement would be to achieve a situation which would give full weight to the obvious merits of UTM, but would temper our decisions with a leaven of common sense in its application. It has been difficult to reconcile the view-points of those concerned and responsible, but I have entertained high hopes.
>
> Lately, however, a senior U.S. official used the following words: 'The United States General Staff have decided that the UTM shall be world-wide and from Pole to Pole, and there can be no discussion on this policy.' Such a statement reduces our chances of agreement almost to the point of invisibility and our future plans to a state of flux, for in no circumstances can I advise the British General Staff to sign such an agreement.

Many other speakers were highly critical of the method of subdividing the world geometrically. This was predictable in Britain because the Greenwich Meridian is the boundary between two UTM zones. Nevertheless the printed discussion does not cover all the negotiations under weigh. Indeed, only a few days later, Baumann and Willis both spoke in support of the motion for adoption of the UTM. Perhaps the clue to this remarkable change in attitude is to be found in two comments made by Hough in winding up the discussion:

- that there was never any intention that the UTM should supersede existing cadastral systems; that it was intended solely for topographical mapping;
- that the Army Map Service were willing to convert the entire South African trigonometric control into UTM coordinates in any form of output that might be required. This invitation appears to have applied generally to all countries.

A clue about the motives for this generosity might be found in the apparently innocent contribution to the discussion by Brazier, who deliberately misquoted Charles Kingsley to indicate that even in those early days of missile diplomacy, 'Happy is the country that has no geodetic datum fixable with reference to UTM'. Within a decade all the military mapping of and most of the civilian mapping of western countries had been converted to UTM.

Computing Transverse Mercator formulae

Although the use of projection tables should have been completely swept away in the revolution of computing methods during the 1970s and 1980s, it is still desirable to know of their existence. Indeed the use of tables to plot the individual terms against the argument of latitude, which was described in the first edition of the book, is still a useful and instructive exercise.

The most convenient of these tables were those prepared by the US Army Map Service (AMS, 1950). Sets of tables for each of the five Figures

of the Earth were published, each in three volumes:

Vol. I: *Transformation of Coordinates from Geographic to Grid*
Vol. II: *Transformation of Coordinates from Grid to Geographic*
Vol. III: *Coordinates for 5-minute Intersections.*

The third volume is intended for cartographic work, because the UTM grid values to plot any graticule intersection likely to be required for small- or medium-scale topographical maps can be extracted from these tables without any computations. Volumes I and II allow the user to make any of the three basic kinds of computation which may be required – determination of position, convergence of meridians and local scale-factor. Another set of tables (DMS, 1958) were prepared in Britain by the Directorate of Military Survey of the Ministry of Defence. These contain the values for all five spheroids in a single slim volume, which were called 'nutshell' tables. Twenty years later, Ministry of Defence (1978) charitably describes them as being 'complex to use and are little employed'. The principal reason for their unpopularity was that, in order to save space, the values of the terms were tabulated at intervals of 20′ of latitude and therefore much interpolation was required. Nowadays, of course, all the necessary computations may be carried out from the original equations by microcomputer, or even by pocket calculator, so that the tables are virtually obsolete.

The solutions using tables had the additional requirement that input angles of latitude and longitude were normally in degree measure, whereas the solutions of the equations were done with the angles expressed in radians. Similarly the various calculated angles were needed in degree measure, and therefore had to be converted from answers expressed in radians. For survey purposes, where the distances from the central meridian are small and convergence does not depart much from zero within the 3° of longitude on either side of the central meridian, these angles are conveniently expressed in seconds of arc. Consequently the conversion from radians to seconds and vice-versa using the sin 1″ convention, described on p. 362, looms large in the UTM and OS equations. However, working in seconds of arc creates other difficulties. The maximum difference in longitude from the central meridian of an UTM zone is 10 800″, which, when raised through several powers, become a very large number. This problem is even more greater in the inverse computations where high powers of E are required. The effect upon the inverse solution is particularly important nowadays because E, in metres, raised to a power such as E^6 may be large enough to trigger off overflow conditions in the computer. For BASIC programs this is at about $1{\cdot}7 \times 10^{38}$, which may be exceeded for E^4, E^5 or E^6 depending upon the distance of the point from the central meridian. Rather greater flexibility is to be obtained if the program is written in PASCAL, for which a maximum capacity of

about 10^{63} is possible. However, it is still quite easy to exceed this in the higher-powered terms of the grid-to-geographicals solutions.

Using tables the normal practice was to work in units of 1/10000 second, so that the maximum difference in longitude from the central meridian is a value of only 1·08. In UTM notation the longitude term is therefore

$$P = \lambda'' \times 10^{-4} \tag{16.46}$$

and, for the inverse solution, the Eastings term is

$$Q = (E - 500000) \times 10^{-6} \tag{16.47}$$

The first two terms of the expression to find E from φ and λ which was given in radians in equation (III.2) now becomes

$$E = 500000 + P(IV) - P^3(V) \ldots \tag{16.48}$$

where

$$(IV) = \nu \sin 1'' . \sin \varphi . \cos \varphi \times 10^8 \tag{16.49}$$

and

$$(V) = (\nu/6) . \sin^3 1'' \cos^3 \varphi (\nu/\rho - \tan^2 \varphi) \times 10^{12} \tag{16.50}$$

Many early computer programs to make Transverse Mercator calculations were not written in a suitably economical form through lack of understanding of the significance of these terms. We know that the sin 1″ trick is intended to transform radians into seconds of arc or vice-versa. We also know that, in order to access the trigonometric subroutines in most computer languages, the angles must be expressed in radians. Consequently it is no longer necessary to make all of the conversions needed when trigonometric functions had to be extracted from tables using the argument of degree measure. However this was appreciated by some programmers, who slavishly followed the presentation of the equations as in Ordnance Survey (1950). It follows that those programs containing all the sin 1″ terms spent a certain amount of CPU time converting angles to and from radians and seconds of arc without furthering progress in the calculations.

A further important factor is the design of the equations. We have already referred to this, together with Vincenty's comments, on pp. 177–178. The nested form of equation for calculation of the meridional arc distance is given in Vincenty (1971), and redesign of the Transverse Mercator projection equations is equally valuable. Not only are they easier to write in an appropriate programming language, they depend upon progressively raising terms to higher powers, and therefore reduce the risk of overflow conditions which are all too common in a full frontal

attack upon Redfearn's equations in Appendix III. We list there the expressions published by Meade (1987) and intended for the UTM, but we have modified them to allow solutions for any scale factor k_0 on the central meridian. These are also listed in Appendix III for direct comparison.

CHAPTER 17

Photogrammetry and remote sensing using conventional photography

> Remote sensing is basically the question of *what* is on the ground, whereas photogrammetry assumes that the operator knows what he/she is looking at and is concerned with *where* it is. However satellite imagery is today being used to make topographic maps. This is admittedly a recent occurrence... and for rather limited scales and purposes, but the fact remains. As traditional mapping becomes more expensive, the demand for new and revised mapping increases and as satellite sensor technology improves, this trend cannot but continue. This does not mean that photogrammetry will become engulfed by remote sensing, merely that we, as photogrammetrists, are making use of a new source of small scale imagery.
>
> D. J. Gugan, *Photogrammetric Record*, 1989

Introduction

It is now desirable to extend discussion about the use of map projections in surveying and mapping to a consideration of their role in the methods of recording and plotting map detail. Since World War II this has almost invariably been done using aerial photography, and usually the final product has been a paper map. We therefore consider first this photogrammetric application, for it has been the bread-and-butter method of spatial data collection for mapping purposes for more than 60 years. Although the normal product from the photogrammetric process has been the line map, alternative forms of reproduction have been various forms of photomaps ranging in sophistication from simple uncontrolled mosaics through orthophotos, which are the most sophisticated form of analogue presentation of the data. During the past 20 years wholly different methods of data collection and processing have arisen through digital methods of storage and data manipulation. The introduction of *analytical plotters*, which are much more flexible in their uses than the *analogue* plotters, has altered the methods of mapping from conventional aerial photographs. For example, the analogue instruments can only make use of the principles of *central perspective*, which is fundamental to conventional photography but is only one of the forms of image geometry

created by sensors. Many of the analogue stereoplotters which were in use a few years ago could only be used with photography of one format, of one focal length or with glass rather than film diapositives. These design restrictions were often unimportant in practice because the type of photography used, and the nature of the work carried out by an agency, corresponded to that for which the plotter had been designed. The instruments were highly efficient for plotting map manuscripts at certain scales but they were inflexible inasmuch as a change in the required output could not be accommodated. Digital processing, which is the basis of the analytical instrument, has created entirely new concepts about the storage and use of positional data. Although an analytical plotter can be used to plot line maps in pencil as formerly, this does not fully exploit the range of the instrument. A most important use of the analytical plotter is to produce digital data for input to a database and hence to GIS applications.

Although photographs of the earth had been taken from space more than a decade earlier, the launching of an earth-sensing satellite by NASA in 1972 led to a new era of mapping on a continuous basis from space. This satellite, called ERTS-1 and renamed Landsat-1 in 1975, was followed by two others, all of which circled the earth in a nearly circular orbit inclined about 99° to the equator and scanning a swath 100 nautical miles wide (185 km) from an altitude of about 919 km. The fourth and fifth Landsat satellites involved circular orbits inclined about 98° and scanning from an altitude of about 705 km. The early Landsat satellite carried two main sensing devices, a television camera (*return beam vidicon*, or RBV) and the *multispectral scanner* (MSS). We shall see in Chapter 18 that the MSS was by far the more important source of data, so much so that the Landsat-4 and Landsat-5 systems dispensed with the RBV, replacing it by a more sophisticated *Thematic Mapper* (*TM*)* which is another multispectral scanning system.

Survey quality photographs have now been taken from manned space vehicles, notably from Spacelab and the Space Shuttle, providing us with material having identical geometry but at much smaller scales than normal aerial photographs. At the time of writing, however, and as a result of the Space Shuttle disaster of January 1986, which ended manned spaceflights from the USA for several years, there are still very few survey photographs taken from space compared with the huge amount of other data which have now been collected by scanning sensor.

* The reader should be aware of possible confusion in contemporary literature between TM standing for Thematic Mapper and TM meaning Transverse Mercator, as in UTM. The author has attempted to avoid ambituity by using the letters TM by themselves to mean Thematic Mapper. The only abbreviation of Transverse Mercator occurs in the initials UTM.

Although they may all look similar, geometrically speaking, aerial photographs are not maps; scanned images are not the same as aerial photographs and they, too, are not maps. A photograph has a central perspective projection as illustrated in Fig. 17.01(a), which is different from the orthogonal projection of a map illustrated in Fig. 17.01(b). This, in turn, differs from the more complicated geometry of the scanning systems to be summarised in Chapter 18.

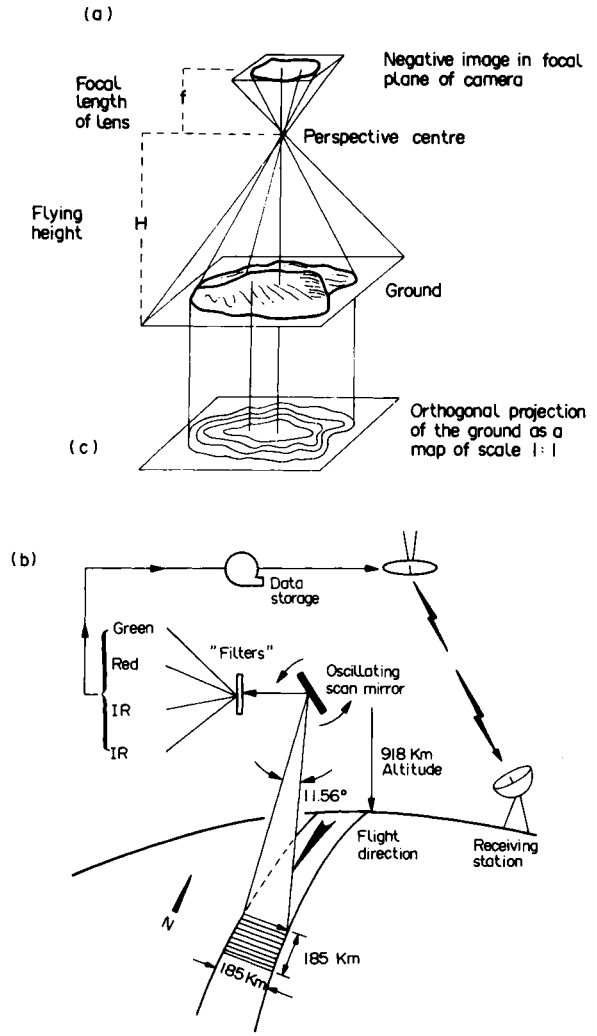


FIG. 17.01 The geometry of the photographs, scanner and map: (a) the central perspective of the simple photographic image; (b) the geometry of image information by a scanning sensor of the form used in Landsat MSS and TM systems; (c) the orthogonal geometry of the map or chart.

Aerial photography and photogrammetry

It is not the purpose of this book to produce a comprehensive introduction to the methods of photogrammetry. That is well provided, for example, by Burnside (1985) and, with special reference to the coordinate systems which are used, by Methley (1986). It will therefore suffice to list the following features of conventional photogrammetric mapping.

(1) The fundamental principle of the geometry of the photograph is that of central perspective; that light rays reflected from a distant object pass through a *perspective centre* as it enters the camera, so that the photograph is formed by a bundle of such rays all converging at the same point, S in Figs 17.01 and 17.02. Moreover, each ray satisfies the condition that object, perspective centre and image are collinear; that all three must lie in the same straight line. A geometrically rigorous method of

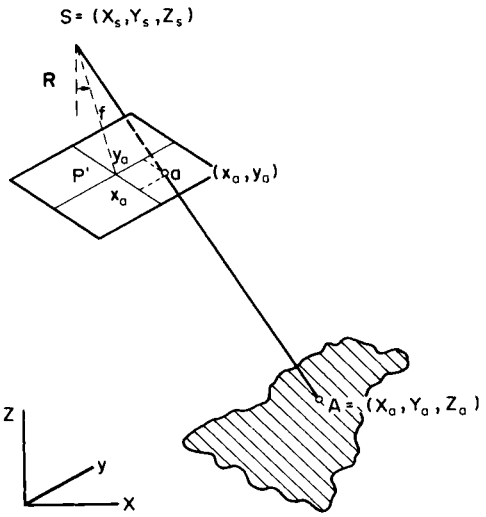


FIG. 17.02 The geometry of collinearity in conventional photograph. An object A is represented upon a photograph by the image a. The perspective centre in the camera lens is the point S. In this figure we employ the construction that S lies at distance f beyond the plane of the photograph, whereas in Fig. 17.01(a) the perspective centre was situated between the ground and the photograph. f is the focal length of the lens. This is the diagrammatic representation of a positive print or diapositive in which images are correctly positioned, whereas Fig. 17.01(a) showed the geometry of the photographic negative in which the relative positions of ground objects are reversed in the film. The collinearity of rays is demonstrated by the straight line SaA. The image coordinates of a are (x_a, y_a) measured from the principal point, P'. Then, if the position of $S = (X_s, Y_s, Z_s)$ and we know the amount of tilt of the photograph, indicated by the rotation matrix **R**, the ground coordinates of $A = (X_a, Y_a, Z_a)$ may be determined.

photogrammetric restitution must reconstruct this bundle of rays at a suitably smaller scale in the plotter.

(2) The instruments and methods for photogrammetric mapping have evolved in favour of using overlapping vertical photographs. By taking vertical rather than oblique photographs the geometry of mapping is simplified. Also the loss of images by dead ground is reduced. By exploiting the principles of stereoscopy through the simple medium of taking strips of overlapping photographs, the third dimension can be recreated and ground height measured.

(3) However, an important restriction upon the use of aerial photography for mapping and measurement is that even when the camera is pointed downwards for vertical photography it is subjected to displacements of the aircraft such as tilting and small changes in flying height. These are notoriously difficult to measure directly with sufficient accuracy from the aircraft. Moreover, in the days before GPS it was not possible to locate the position of the aircraft with sufficient accuracy to dispense with other sources of information. Consequently most photogrammetric mapping methods are *indirect* in the sense that tilts and other displacements have to be measured and corrected by examination of the photographic images themselves, and the photographs have to be oriented to the earth's surface with reference to surface features, or *ground control* which has been fixed independently.

(4) An image point may be referred to the *principal point* or geometrical centre of the photograph by means of (x, y) *image* or *plate coordinates*. Each image may be defined thus with reference to the centres of two photographs taken from different places. It is necessary to relate them to one another by reconstructing the conditions which occurred during photography. This involves conversion from the plane (x, y) coordinates into the three-dimensional (X, Y, Z) system within the plotter. We refer to this three-dimensional reconstruction as the *stereoscopic model*, *stereomodel* or more simply, *model*.

(5) The relationship between photographic image and plotting pencil is provided by a suitable optical system within the instrument so that the operator may observe the stereoscopic image of the model as outlines are traced. A measuring mark is needed to relate photographic image to plotting mechanism. Because a pair of photographs are viewed stereoscopically there need to be two marks within the optics; one for each photograph of a pair. These also appear to fuse stereoscopically, thereby creating the illusion of a *floating mark*. Movements of these marks about the planes of the photographs connect through various linkages to control the drives of the plotter and ultimately measure the model coordinates or plot the detail to be mapped. The third dimension, Z, is, of course, ground height. This is measured by observing the floating mark as it just appears to touch the ground surface, when the Z coordinate may be

read from a suitably graduated height scale. In analogue plotters the X, Y and Z axes are represented by rigid steel bars.

(6) Creation of the stereomodel involves the elimination of the small differential tilts which have occurred between the instants when the two photographs were taken. This is known as *relative orientation.* Comparison of points in the resulting model with the plotted positions and known heights of the control points allows alteration of model scale and tilting of it until the datum surface in the model corresponds to that of the ground. This is known as *absolute orientation.*

(7) In order to relate the plotting sheet to the photogrammetric model, it is placed upon the drawing table or a coordinatograph connected to the instrument and shifted around until the images of the ground control points viewed with the floating mark coincide with the positions of the pencil point above their plotted positions. The scale of the model is adjusted for any discrepancy in distance between two ground control points viewed with the floating mark placed upon each in turn, and the corresponding distance plotted to scale on a map. By means of successive approximations the comparison is repeated until the instrument can plot at the required map scale.

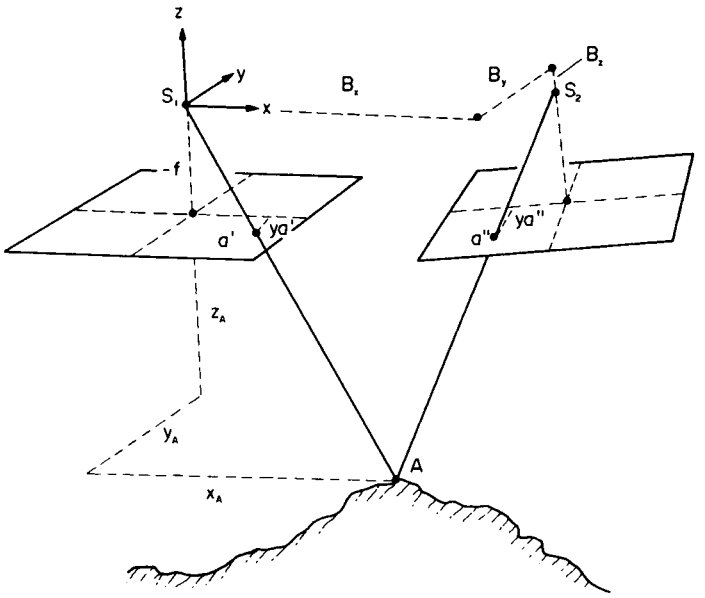


FIG. 17.03 The three-dimensional model created by a stereoscopic pair of photographs set in an analogue plotter showing the plate coordinates measured to determine the parallaxes y'_a and y''_a of a′ and a″ which are the corresponding images of the point *A*.

If a numerical solution is used, the (X, Y) coordinates are related to their corresponding ground positions (E, N) through the grid-on-grid transformations described in Chapter 2. This may be done through the simple conformal transformation; indeed this is the only possible method if the minimum of only two points are known, but affine and higher-order transformations may be used if additional ground control has been established.

(8) It follows that the XY surface within the plotter, which serves as the datum for height measurement as well as the surface upon which the model is reconstructed, approximates to a plane. Hence photogrammetric restitution in most analogue plotters is based upon the plane assumption, and generally no attempt is made to correct for the effects of earth curvature. On the other hand suitable numerical corrections may be created in analytical plotters.

The geometry of the photogrammetric plot

The ground control used for absolute orientation has normally been computed on the projection of the national survey. Additional points surveyed specifically for the work are often fitted to the national coordinate system using plane survey methods. At the scales usually employed in photogrammetric mapping and the corresponding amount of the earth's surface which can be accommodated on the plotting sheet, any difference between a point plotted on the national version of the Transverse Mercator projection and that derived from a local plane survey referred to existing survey control points is less than the zero dimension. Within the single model the influence of earth curvature is also very small. In most conventional cartographic applications we have to accept a discrepancy between the plotting pencil and the observed point which is as large as the zero dimension, simply because in making adjustments by successive approximations the stage is reached where no further reduction of error is possible; where any further adjustment of the instrument will simply introduce a larger error in the opposite direction. Similarly planimetric discrepancies between plots of separate models must also be tolerated if they are no greater than the zero dimension. It follows that fitting a series of pencil plots to control points plotted on the required projection involves a similar amount of uncertainty in location. On the paper map this can be tolerated, but in the output of photogrammetric data in digital form, a zero dimension of this magnitude is no longer acceptable and it is necessary to work to finer tolerances. In fitting the detail to a projection this means that 'near enough' is no longer 'good enough', and that it may well be necessary to compute the projection coordinates of points of detail using the higher-order terms which were conveniently ignored in ordinary plane surveying.

Ground control and its identification

Because of the need for ground control in absolute orientation an important stage in the mapping process is the establishment and identification of ground control points. For topographical mapping the location of control within the national or state grid system is done by field survey methods. It is relatively uncommon to use the positions of conspicuous objects already shown on a map for this purpose. In new topographical mapping this is simply not possible, because a suitably reliable map does not exist until the photogrammetric work has been completed. However, even in a well-mapped area uncertainties arise from using map detail simply because we cannot regard the positions of control points which have been obtained by measurement on a map as being without error. There is bound to be some uncertainty about the accuracy of the cartometric measurements, as well as the possibility of there being positional errors in the map itself.

The distance calculated from the coordinates of two points determined by field survey methods also contains some errors, for no measurement process is every wholly free from them. However, survey measurements are made on the ground, and positions of surveyed points and the required distance between a pair of points are computed in metres on the ground without reference to any map scale. Consequently any residual errors in the survey are very much smaller than the zero dimension of any map to be made from it. Therefore a distance calculated from field observations is preferable to a map distance measured by ruler and then converted into suitable units. Because determination of ground control by field survey methods is slow compared with the time needed to take the photographs, it is usual for production bottlenecks to occur at this stage. It also follows that provision of a network of ground points to control each stereomodel individually leads to an uneconomically high density of survey stations, because the number of control points is related to the number of photographs, not to the area of the block of country to be mapped. See Maling, in Goodier (1971), for a discussion of this subject. Consequently special photogrammetric methods have been evolved to create supplementary control of the overlapping photographs forming a *strip* or even a *block* of photography. These are the methods of *aerial triangulation* which have become immensely important in conventional photogrammetry during the past 50 years. Theoretically it is possible to orientate a strip or even a block of photography with the same amount of ground control as is required to orientate the single photograph, but it would be most unwise to rely entirely upon so little control information, because many of the detailed operations involved in aerial triangulation may introduce systematic and therefore cumulative errors into the work.

The influence of earth curvature on aerial photography

In the foregoing description of absolute orientation it is implicitly assumed that the earth is flat. However, we have already referred to a photograph as being a central projection of the ground. This is a perspective azimuthal projection with its origin at the nadir point. In Fig. 17.04(a), N is the *nadir point* on the earth's surface directly beneath the camera and n is its homologue on the plane of the negative. Through the nadir point is a plane, represented in Fig. 17.04(a) by the line NA', tangential to the earth at N. Then a point A on the earth will be depicted on this plane at the point A' where the ray SA intersects the tangent plane. In effect, therefore, the representation NA' is a perspective azimuthal projection of the earth's surface. In Fig. 17.04, SN is the flying height of the aircraft, H; ON is the radius of the earth; and the sum of these distances, SO = D. We denote Δ = D/R, from which it follows that

$$\Delta = (H/R)+1 \tag{17.01}$$

$$H = R(\Delta+1) \tag{17.02}$$

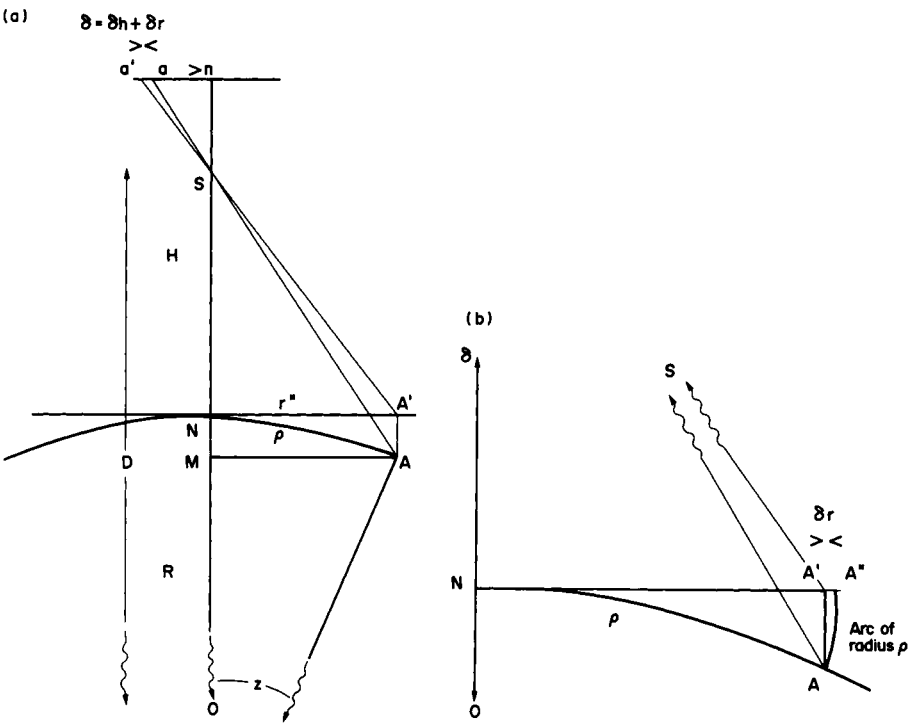


FIG. 17.04 The geometry of earth curvature in conventional aerial photography, showing: (a), the principal effect of curvature upon measurement of height; (b), the much smaller planimetric displacement.

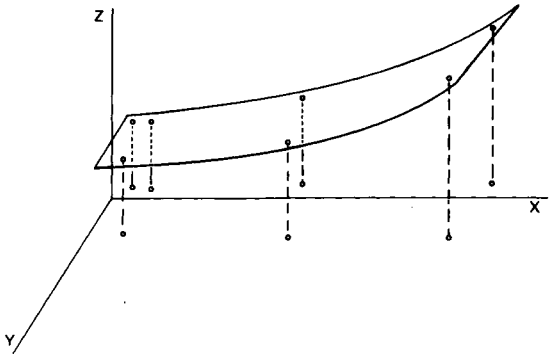


FIG. 17.05 The plane assumption in aerial triangulation. The deformed surface represents a strip of aerial photographs in which common points on adjacent photographs have been linked to one another by measurement of their X, Y and Z coordinates in a photogrammetric plotter. Because each stereoscopic model is treated as if this were part of a 'flat earth' and height Z has been measured from the XY plane, the influence of earth curvature is to produce a systematic increase in apparent height as the triangulation proceeds. Thus the influence of earth curvature is reflected by the apparent curvature of the strip (indicating increasing errors in height) in the opposite direction. It should be noted that other kinds of systematic error may also occur, so that the upward warping may be modified. However, the dominant cause of the upward warp is earth curvature and this is always present in triangulation of a long strip of photographs.

We may now determine the radius vector $\rho = NA'$

$$\rho = \mathrm{R}(\Delta - 1) \sin \mathrm{z}/(\Delta - \cos \mathrm{z}) \qquad (17.03)$$

In addition a spherical angle, θ, in the plane NA' orientates the line NA' with respect to some datum direction such as true north. Then the position of A' may be referred to the origin N by the plane polar coordinates (ρ, θ). Figure 17.04(a) only shows the influence of earth curvature in section. In the tangent plane we are comparing the (ρ, θ) coordinates and the (z, α) bearing and distance coordinates of the point A on the spherical surface. Because we are dealing with a tangent plane, the spherical angle, α, is equal to the plane angle θ, as is contained in the definition of a spherical angle (p. 54) and the fundamental properties of all azimuthal projections (p. 128). It follows that the only displacements of significance are in a radial direction from the point of tangency. In Fig. 17.04(b) the image displacement which occurs on the plane of the negative is shown. Because of the earth's curvature the image point corresponding to A is a and that corresponding to A' is a'. We denote the distance pp' on the negative by δ and the distance $na = r$.

The first component is the planimetric shift in the position of A to A' in the horizontal plane; the second component is the *height displacement* from A to A'. A cursory glance at this diagram indicates that this dis-

placement, δ_h is much greater than the planimetric displacement δ_r, but, because of the geometry of image formation, both operate in the same direction on the photograph. Thus the displacement a–a' on the negative plane is $\delta = \delta_h + \delta_r$.

It should be clear that, on a typical aerial photograph, the maximum distance ρ and therefore the angle z is very small. On a photograph having the 228 mm × 228 mm format, which is the standard negative size used for survey cameras in English-speaking countries, the maximum distance from the nadir point is one-half of the diagonal (171 mm) to the corners of the frame. Within the range of scales which are commonly obtained from the survey aircraft used for civilian purposes, this distance is at most 14 km, so that the most remote point on a photograph is not far from the origin of the projection. Therefore we may forecast that the displacements are also small.

The planimetric component of earth curvature

Figure 17.04(b) illustrates the influence of earth curvature as this affects the distances NA and NA'. On the tangent plane the planimetric component is

$$\delta_r = R \sin z - Rz \tag{17.04}$$

$$= R(z - z^3/6 + \ldots) \tag{17.05}$$

Neglecting higher-order terms we may write

$$\delta_r = -Rz^3/6 \tag{17.06}$$

and because $R \,.\, z \approx \rho$,

$$\delta_r = -\rho^3/6R^2 \tag{17.07}$$

In the example of a maximum distance of 14 km from the origin of the projection, $\delta_r = 0{\cdot}01$ m. For $\rho = 100$ km, δ_r is about 5 m or only 1 part in 20 000. For most conventional cartographic purposes this is small enough to be ignored.

The altimetric component of earth curvature

Because the planimetric component is so small, it is ignored in the construction of Fig. 17.04(a) and the influence of curvature upon height measurements is indicated by the distance PP'. From the triangle NAA'

$$AA' = R - (R^2 - \rho^2)^{1/2} \tag{17.08}$$

$$= R[1 - (1 + \rho^2/R^2)^{1/2}] \tag{17.09}$$

or

$$\delta_h = \rho^2/2R \tag{17.10}$$

Putting R = 6371 km, an approximate version of this correction is

$$\delta_h = 0{\cdot}0785\,\rho^2 \tag{17.11}$$

where δ_h is expressed in metres and ρ in kilometres. Thus for $\rho = 100$ km, $\delta_h = 785$ m, which demonstrates the big difference between δ_h and δ_r. At the scales of photography used for topographical mapping (1/10 000–1/100 000), the half-diagonal distances represent very much shorter distances on the earth so that the corresponding effects upon heights are also much smaller. The displacement is 0·2 m at scale 1/10 000, 1·3 m at 1/25 000 and 20·3 m at 1/100 000. The values relate to the tangent plane. On the film itself the image displacement $p'p''$ may be determined from

$$\delta = Hr^3/2Rf^2 \tag{17.12}$$

where $r = na$, f is the focal length of the camera lens (both expressed in millimetres), H is the flying height in the same units as earth radius, R.

An important source of systematic error in photogrammetry which is of immediate interest to us is the effect of earth curvature upon strip triangulation. The object of triangulation, as we have seen, is to provide control in those parts of a block of photography for which no ground control is available. The classic procedure is to create each model in an analogue instrument, using special techniques to join each model to that preceding it. Because analogue restitution almost invariably means employing the plane assumption for each model, the formation of a strip means joining together a succession of models each of which has been reconstructed within the XY plane of the plotter. The result of forming the strip with reference to the XY plane, whereas the photographs themselves were taken of the curved surface of the earth, leads to increasing divergence between the plane and curved surfaces expressed as systematic height errors. This is illustrated by Fig. 17.05, which shows the characteristic deformation experienced in strip triangulation. This shows that as the distance X increases from left to right along the strip, the measured heights exhibit increasingly large errors, but because these are measured from the plane surface represented by the base carriage of the analogue plotter, these negative errors simulate a warping of the strip in the upward direction. From (17.08) the height error (ΔZ) can be expressed by the equation

$$\Delta Z = R - (R^2 - X^2)^{1/2} \tag{17.13}$$

where R is the earth's radius and X, Z are coordinates within the system,

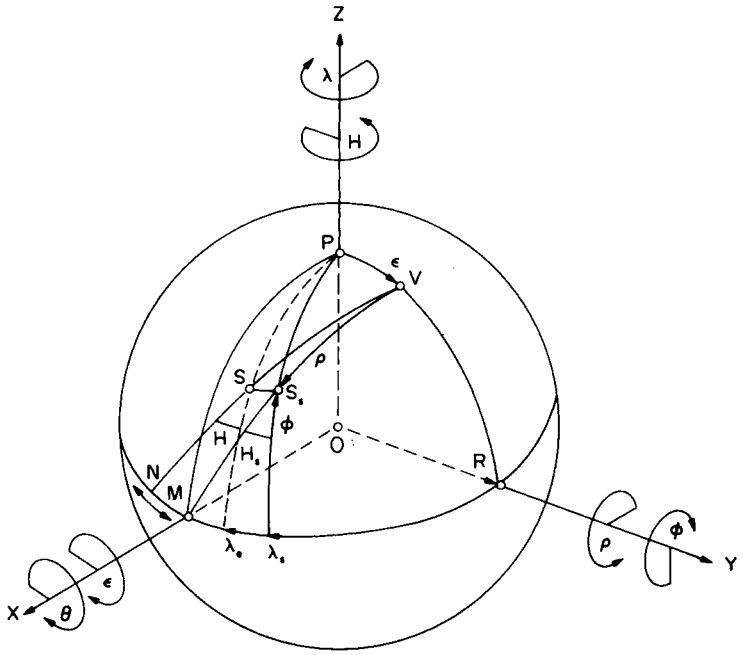


FIG. 17.06 The orbital geometry of a satellite. (Source: Kratky, 1974.)

illustrated in Fig. 17.06. Finally, as in equation (17.10)

$$\Delta Z = -X^2/2R \qquad (17.14)$$

We may demonstrate the magnitude of this from the following example. Suppose that we have a strip of large-scale vertical aerial photography, for example at 1/10 000, and we carry out strip triangulation through 10 overlapping models. The total length of the strip will be of the order of 11·5 km, and even assuming that there are no other sources of error in the strip, the amount of upwarping will amount to a negative height error in the final model of 10 m.

Photographs taken from artificial satellites

Table 17.01 lists the few occasions when photographs of survey quality have been taken from artificial satellites. The interruption to the US effort after 1986 is obvious. Indeed, the ill-fated Challenger was equipped to take the next series of Space Shuttle photographs. The Soviet contribution in the later years has been considerable, but this has only recently become available to Western users. Consequently it is still too early to make any proper comparative evaluation.

The obvious difference between conventional photography taken from

TABLE 17.01 *A summary of survey quality photography from space*

Sortie	Year of sortie	Country	Camera	Scale	Ground resolution (metres)
Skylab	1974	USA	S190B	1/950 000	40
Skylab	1974	USA	S190A	1/2 850 000	100
Soyuz	1976	USSR	MKF-6	1/2 000 000	20
Space Shuttle (41-G)	1982	USA	LFC	1/800 000	10
Spacelab (Metric Camera Project)	1982	Europe	RMK 30/23	1/820 000	20

an aircraft and that taken from a satellite is scale. Whereas the smallest scale of photography which can be obtained using the types of aircraft commonly used for civilian survey photography is within the range 1/70 000 to 1/100 000, the scale of photographs taken from space is considerably smaller. The first survey-quality photography was taken on the American Space Shuttle Mission 41-G of 1982 using the *LFC* or *large format camera*, and the European Space Agency *Spacelab-I* mission of the following year which carried a standard Zeiss 30/23 Metric Camera. These sorties produced aerial photographs of scale 1/800 000 and 1/820 000, respectively. See Meneguette (1985) for an account of orientation of the photographs taken during the Metric Camera Project.

Return beam vidicon (RBV)

The return beam vidicon camera mounted on Landsat-1 through Landsat-3 was a television camera. It created an image which was also to all intents and purposes, instantaneous. It follows that the geometry of the RBV image for mapping purposes is virtually the same as that for conventional photography. Since the format of the RBV is 185 km × 185 km, the maximum value for ρ for the half-diagonal has only increased to 132 km. Consequently from (17.11) and (17.14) the planimetric component δ_r is still only 9·4 m, but the height correction δ_h has now increased to 1366 m. Compared with the huge success of the multispectral scanning systems (MSS) on all Landsat sorties, the RBV had a somewhat unfortunate history, for the original Landsat-1 instrument failed quite soon after launching, and the only really useful cover was obtained from Landsat-2 and Landsat-3.

Orbital geometry

An essential preliminary to any consideration of the use of map projections for any kinds of photographs or scanned images taken from space is the recognition of two separate variables.

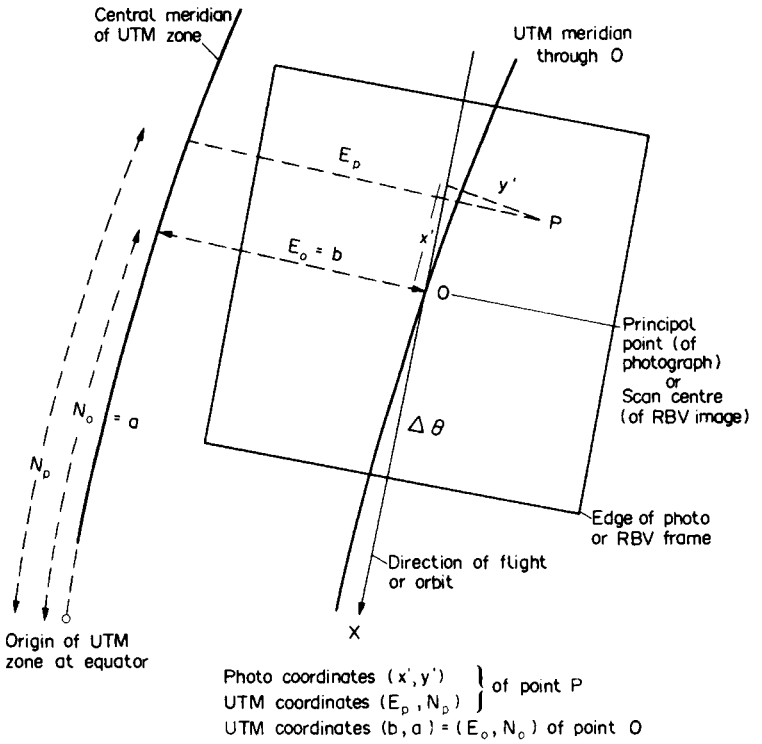


FIG. 17.07 The transformation from RBV or aerial camera coordinates (x', y') into UTM coordinates (E, N). (Source: Kratky, 1974.)

- *Earth curvature* effects are those which have been described above, and which also affect larger-scale aerial photographs, but obviously to a lesser extent. The displacement δ_r is the exact equivalent to linear distortion in an azimuthal projection described in preceding chapters.
- *Earth rotation* effects are peculiar to satellite imagery and are the result of the earth's rotation while the satellite is orbiting above it. There are two such effects. The first of these influences location of an image of the earth with respect to the orbit of the satellite. Although the orbits of sun stationary satellites used for collecting terrestrial data are often nearly circular, and it seems that the orbital geometry is quite simple, so that the arc ρ may be calculated without significant error by spherical trigonometry, the simple idea that the point S_s in Fig. 17.06 represents the position of the satellite with respect to the earth at the time of exposure is not valid. Because the earth has rotated independently beneath the arc traced by the orbit the satellite *appears* to occupy the position S. It follows that we must compute revised geographical coordinates (φ, λ) for the sub-satellite

point beneath S and not the simple solution of a spherical triangle to find S_s.

The second effect of earth rotation is skewing of the scanned images produced, for example, by the Landsat MSS, the TM and the HRV images obtained from the SPOT satellite. This results from the time required to collect information for a whole scene by a scanner which operates one line at a time. During creation of the image the satellite is proceeding along its orbit and the earth is rotating beneath. The effects are described in Chapter 18.

Orientation of satellite photography involves location of the satellite and its nadir, or *sub-satellite point* with reference to the earth. This is done using the orbital characteristics of the particular satellite. In Fig. 17.06, V is the vertex of the orbit and is taken as the origin from which to measure orbital travel distance, ρ, and also for determining the reference meridian against which we measure the change in longitude λ_s. Assuming that the earth is stationary and not rotating, the nominal position of the satellite at a particular instant of time is S_s, which may be located by the orbital parameters (ε, ρ) or by the corresponding change in geographical coordinates (χ, λ_s) where χ is the colatitude of the sub-satellite point PS_s. The nominal heading of the satellite towards the local meridian is the angle θ_s.

If the orbit of a satellite is circular, the surface containing the points P, V, S_s and S in Fig. 17.06 is a sphere of radius equal to the distance $OP = OS = H + R$, composed of the radius of the earth plus the height of the satellite above the earth. We already know that the earth is not a perfect sphere; moreover, the orbit of the satellite may also be elliptical. Therefore neither H nor R is constant. However, we proceed from the simpler assumption that they are.

The rotation of the earth affects the actual position of the sub-satellite point, displacing it additionally in the direction of the parallel φ from S_s to S. It follows that the value of φ is unaltered, but λ is increased by an amount λ_e and the sub-satellite track gradually deviates by the angle $\Delta\theta$ from the heading θ_s. These changes are proportional to r_e, which is the ratio between the rates of change in position of satellite and earth. For the early Landsat vehicles the satellite completed full earth cover in 18 days, and after 251 orbits. Therefore, $r_s = 18/251 = 0{\cdot}071\,7\ldots$ and is constant for any particular satellite. For early Landsat, $\varepsilon = 9°{\cdot}092$. Assuming both a spherical earth and circular orbit, the geographical coordinates of the nadir or sub-satellite point N and the real heading at this point may be expressed by the following equations

$$\varphi = \sin^{-1}(\cos\varepsilon \,.\, \cos\rho) \tag{17.15}$$

$$\lambda = \tan^{-1}(\tan\rho/\sin\varepsilon) + r_s \,.\, \rho \tag{17.16}$$

$$\theta = \tan^{-1}(\tan \varepsilon / \sin \rho) + \tan^{-1}(r_s \cos \varepsilon \sin \rho) \qquad (17.17)$$

A truly vertical aerial photograph taken from the satellite positioned at S is centred upon the nadir, N, and the ground track is taken to be a great circle arc oriented on the bearing $180° + H_s$. The definition corresponds to a transverse cylindrical projection of the form of rectangular spherical (Cassini) coordinates as illustrated by Fig. 15.04. Therefore it is easy enough to express position on a very small-scale aerial photograph (or RBV image) in the form of Fig. 17.07, where a point A on a photograph (or RBV image) having plate coordinates (x′, y′) to the centre of the scene, which in a conventional photograph is the principal point and in any image from which tilt is absent is the sub-satellite or nadir point.

$$\begin{aligned} N_p &= N_0 + u + u(b+v)^2/2R^2 \\ E_p &= E_0 + v - u^2 b/2R^2 + (b+v)^3/6R^2 \end{aligned} \qquad (17.18)$$

where (E_0, N_0) are the grid coordinates of the principal point, R is the radius of the spherical earth and (u, v) are the plate coordinates rotated through the angle $\Delta\theta$ which is the heading of the satellite when it is situated at $0 = (\varphi_0, \lambda_0) = (E_0, N_0)$.

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos \Delta\theta & -\sin \Delta\theta \\ \sin \Delta\theta & \cos \Delta\theta \end{pmatrix} \cdot \begin{pmatrix} x' \\ y' \end{pmatrix} \qquad (17.19)$$

There are consequently three stages in determining the UTM coordinates of a point on a photograph.

1. The geographical coordinates (φ_0, λ_0) of the sub-satellite point are calculated using the orbital equations (17.15)–(17.17).
2. These are converted into the corresponding UTM coordinates using the geographicals–grid transformation described in Chapter 16 to find (E_0, N_0).
3. Equations (17.18) are now used to calculate E_p and N_p. Although these expressions are derived for the sphere and not the spheroid, within the format of the single photograph or RBV image the errors resulting from this assumption are less than the zero dimension corresponding to the resolution of the images. Finally the amount of distortion in position determined from the photograph may be expressed as

$$\left.\begin{aligned} dN &= (E_0 R^2) \cos\theta [x'y'(\cos^2\theta - \sin^2\theta) + (x'^2 - y'^2) \sin\theta \cos\theta] \\ dE &= (E_0/R^2) \sin\theta [x'y'(\sin^2\theta - \cos^2\theta) + (y'^2 - x'^2) \sin\theta \cos\theta] \end{aligned}\right\} \qquad (17.20)$$

Kratky has shown that dN and dE reach maximum values at the equator and the edge of a UTM zone, where it amounts to 61 m in the x′ direction and 14 m in the y′ direction. In higher latitudes and, of course, nearer the central meridian of each zone, the errors are appreciably less.

The method outlined could be used to transform data point-by-point, but it is too slow and complicated to use this for the transformation of photographic detail into map detail. We shall see in the next two chapters that other methods must be sought to accomplish this efficiently.

CHAPTER 18

Projection transformations employed in photogrammetry and remote sensing using scanning sensors

> At first sight it seems extraordinary that it should be cheaper to go 500 miles out into space to get this sort of information, rather than sending out someone on a bike to get it on the ground. However, once in orbit a satellite requires virtually no maintenance, and when compared with the tedious process of first obtaining and then collating all the survey and other details required in ground recording, using its products may require less effort and expense.
>
> J. W. Wright, *Land and Minerals Surveying*, 1988

Introduction to scanning systems

Scanned images are recorded on magnetic tape as a stream of data referring to the discrete rectangular cells known as pixels. The data collected are a succession of signals each relating to the location of a pixel and the character of the image, expressed by a number corresponding to the spectral reflectivity of the ground surface in that cell. Two different types of sensor are used:

- The scanner comprises a swinging or rotating mirror which traverses across the scene and collects information as it sweeps from side to side. This is illustrated by Fig. 18.01(a). Because of the motion of the sideways sweep, it is sometimes colloquially called a *whisk-broom scanner*. It is the design of scanner used for both the Landsat MSS and TM systems.
- The second design is the so-called *push-broom scanner*, in which a battery of sensors are aligned in the athwartships direction of the sensor so that each line of information is collected simultaneously as in Fig. 18.01(b). This is the form of scanner used in SPOT HRV.

It follows that the most important difference between the methods of photography and television sensing which were described in Chapter 16 and these methods of data collection is that whereas a photographic

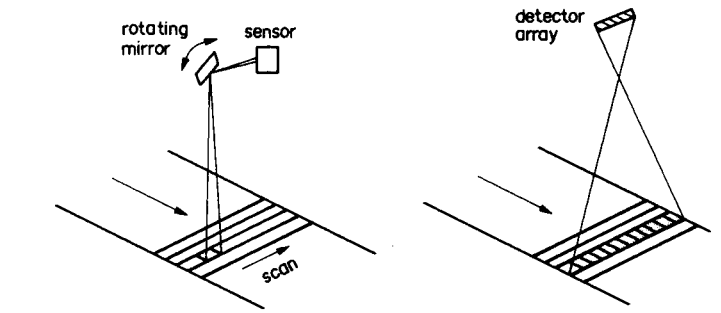


FIG. 18.01 A comparison of the two types of remote sensing scanner (a) the 'whisk-broom' or swinging mirror scanner, as used for the Landsat MSS; (b) the 'push-broom' scanner used for the SPOT HRV scanning system.

exposure is practically instantaneous, scanning is a *dynamic* process, requiring some time to collect all the images comprising a single scene. The scanning time is 28·63 seconds for Landsat MSS, falling to 9 seconds per frame for SPOT HRV. These delays are long enough for image formation to be affected by both the orbital motion of the satellite and earth rotation.

Objectives and limitations

Whereas most conventional aerial photography has been acquired for the primary purpose of map making, and the specialised interpretation of it has generally been a secondary consideration, experience in the use of MSS imagery has been the opposite. This is partly owing to two important limitations to the Landsat products; first the relatively poor resolution (79 m) of the early Landsat MSS so that, as a source for plotting detail, it was suitable only for mapping at the topographical scales smaller than 1/250 000, and it could not compete with conventional aerial photography at the typical mapping scales at 1/50 000 or larger; secondly that MSS images could not be used to measure height and plot contours. With the development of TM and HRV imagery the limits imposed by resolution are much less important. In particular, SPOT imagery is being increasingly used for topographical mapping. However, as Dowman (1985) has indicated, these are still only suitable for medium- or small-scale work. The really important role of MSS imagery has been in the field of specialised interpretation; for example, evaluation of ground cover, land use and crop evaluation. The relatively small scale permits a synoptic view of the variables without having to handle large numbers of individual aerial photographs. The ability to obtain (cloud cover permitting) repeated cover at intervals of only 18 days has further made

multitemporal analysis a far more useful tool than was hitherto possible. Consequently there has been great emphasis in the use of comparative cover of an area, comparing images created at different wavelengths and images created at different seasons. This means that in geometrical terms it is often more desirable to obtain good *registration* between successive pictures than good *rectification* of the image to a map.

Conventional photography has normally been plotted by analogue methods, and until recent years digital processing has played only a small part in photogrammetric mapping. In dealing with scanned data, digital processing is essential. Consequently, in any consideration of the treatment of the digital data forming Landsat MSS scenes, the user's needs must be tempered by knowledge of what is practically feasible.

A terminological diversion

We have just introduced two words to indicate the relationship between different images and maps; *rectification* and *registration*. It is desirable to define these more precisely and introduce two additional words: *resampling* and *rubber-sheeting*, both of which appear in the literature on the subject.

Rectification

The word *rectification* was originally used to describe the methods of correcting conventional photography to fit ground control. The classic way, described in the textbooks of photogrammetry, was to simulate the tilts affecting an aerial photograph by projecting the image of the single aerial photograph to fit the plotted positions of ground control points. This form of rectification is an approximate method of mapping because it is not possible to correct for the displacements owing to surface relief. Because it is part of a photographic process, a mosaic made from a collection of similarly rectified prints makes a satisfactory photomap. The word has been less commonly used as a synonym for restitution, to describe the analogue mapping processes applied to a pair of photographs set in a stereoplotter. By this route, rectification has become one of the words describing the geometric correction of a remotely sensed image. It is used in the absolute sense that the images have been transformed to their map positions established by the use of ground control points as in the absolute orientation of a photogrammetric model.

Registration

The word *registration* comes from the terminology of printing via conventional cartographic practice, where it describes matching two or more

colour separation drawings or printing plates to produce multicoloured images in which each line element or screen dot occupies its correct position relative to the other images in other colours. The action of 'bringing two plates into register' is an extremely skilled manipulation of the position of the printing plate in the press until this is achieved. It is therefore carried over into remote sensing practice to describe the process of matching one set of images to another. Strictly speaking there is no need to rectify a scanned image if, for example, it is only required to compare multitemporal scenes. It will suffice to register the second image to the first, by making some image points correspond to one another without needing to consider the positions of them on the earth's surface. The digital process of handling remotely sensed files usually involves solution of a polynomial expression which distorts one of the images to fit the other.

However, some writers logically regard the process of superimposing the image to ground control as being a form of registration. This usage makes the word rectification redundant, although it misses the important point that satisfactory registration can be carried out without any ground control information, provided we do not want to make a map from the scene.

Resampling

Resampling is the means by which a geometric transformation is actually applied to the input data. This is a more sophisticated procedure brought about by the fact that the pixel coordinates (r, c) of a scanned scene are integer values determined by counting along the scan lines and the number of lines scanned. However, when a scene has been transformed by the various movements of translation, scale change and rotation, the values corresponding to the original pixels are converted into real values, and expressed as decimals. Then it is necessary to recover the pixel locations for the transformed scene without losing vital information or repeating the same pixel more than once. Some of the ways in which resampling may be done are illustrated in Fig. 18.07, p. 399.

Although the term was originally intended to apply to the transformation applied to scanned images in this fairly restricted context, it has now been extended to mean a variety of other transformations and in this way to the manipulation of GIS files, particularly those created by raster digitising, which have no relationship to remotely sensed imagery. Thus Tobler (1988) describes other uses of resampling in making changes from one map projection to another which corresponds to some of the methods to be described later in Chapter 19. They are mentioned here to indicate that there is no really clear-cut definition of resampling to isolate this activity from a variety of other kinds of mathematical transformation.

Rubber-sheeting

The use of this term contains the idea that the original image, or in GIS transformations the source map, may be likened to an elastic sheet which can be stretched in all directions. A network of points in a scene which has been partly corrected in preprocessing is compared with the network of corresponding fixed points (ground control points) and the vectors formed between corresponding point pairs are the mapping deformations. By applying the rubber-sheeting algorithms the original scene is stretched or shrunk until the vectors approach zero, and it may then be assumed that all other pixels on the original scene have been relocated correctly. Rubber-sheeting is done interactively, rather than using an explicit single mapping polynomial such as those described elsewhere in this chapter and in Chapter 19.

Data storage requirements

It is necessary to know the capability of a computer to store and handle the amount of data required, and it is important not to underestimate the magnitude of the task. The following example is instructive.

The amount of Britain covered by a single MSS scene is shown in Fig. 18.02. It can be seen that a total of 55 scenes are required to cover the British Isles entirely, although some of these are predominantly pictures of the sea. The single MSS frame of format 185 km × 185 km comprises 2340 scan lines of 3240 pixels each, or more than $7\frac{1}{2}$ million pixels, and since there are four separate scenes corresponding to the four different wavelength bands used by the MSS, the total information content for each scene is in excess of 30 million pixels. In total, the cover for the British Isles and intervening sea areas therefore comprises more than 1500 million pixels. The Landsat TM comprises 5700 lines of 6900 pixels each, or more than 39 million pixels per scene and since it operates at seven different wavebands, the storage requirement per scene is of the order of 262 Mb. Such large datafiles obviously create considerable problems of storage of the information and, of course, problems of rapid access to any part of it. Moreover, complicated numerical transformations applied to every pixel in turn will make excessive demands upon computing time, so that the more obvious methods of geometric correction may not be the most economical.

Preprocessing or initial system corrections

In the raw form, as broadcast by the satellite, the MSS data need numerous modifications and corrections. For example, the stream of sensed data is collected continuously, so that the first job has to be division of it into

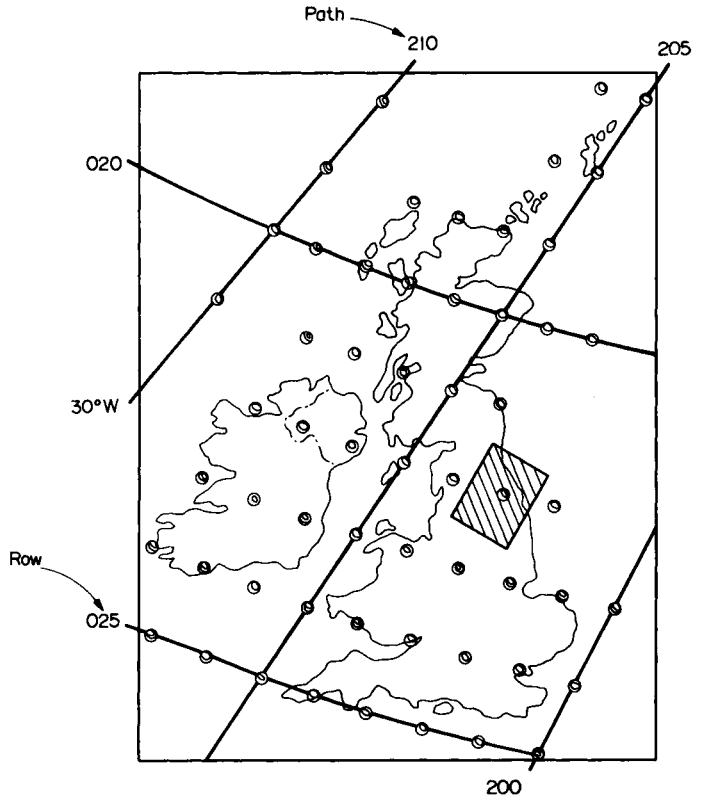


FIG. 18.02 Cover diagram showing the nominal positions of the centres of Landsat MSS images and the extent of the full scene (shaded rectangle) for the later (Landsat-4 and Landsat-5) imagery. (Source: National Remote Sensing Centre.)

blocks of 2340 scan lines, each block representing the separate frame or scene. Some of the flaws and deficiencies of the stream of raw data are *radiometric* and some are *geometric*. The *radiometric* corrections are needed to improve image quality and interpretability, therefore they are mainly cosmetic. Geometric correction of an image involves repositioning the sample elements from where they are to where they should be. A balance must be maintained between special radiometric and geometric processing which depends upon the subsequent use to be made of the images. For example, if it is intended to produce a detailed map of the distribution of vegetation or land-use types, the presence of any residual geometrical distortions in the scenes obviously reduces the accuracy of any maps produced from these data. On the other hand if the purpose of the study is to establish the presence or absence of a particular category of vegetation or land use, a visual comparison will suffice. Because the

study is concerned with determining the presence or absence of a category, rather than its precise location, the geometrical distortions of the image will be of secondary importance. Similarly in making the comparison between different versions of the same scene obtained on different occasions, which is used to detect change, it is sufficient for the different images to be located relative to one another or to be in good registration, but determination of accurate absolute position on the earth is not necessarily needed.

In this chapter we are concerned only with the geometric corrections, and we further narrow this field to the consideration of those needed to compensate for earth curvature and rotation, together with those projections most suitable for presentation of the scanned scenes on a map.

In preprocessing, the location of a scene is normally calculated from the satellite's ephemeris; that is to say, from the assumed position of the satellite in its orbit at the time the image is being sensed. However, this determination may be prone to error. For example, NRSC (1985) indicate that the positions of the nominal scene centres (as shown in Fig. 18.02) can differ from the actual image centres by up to 20 km for Landsat-4 and Landsat-5. The work by Bryant *et al.* (1985), using precision-processed TM images from Landsat-4, demonstrated that it was the indifferent quality of the ephemeris data which had greatest influence upon the accuracy of their work. This is, of course, the same difficulty which we have in conventional photogrammetry, where there has been no satisfactory method of independently locating the position of the aircraft with geodetic accuracy and, as in air survey, this difficulty has to be overcome by referring the images of ground control points to their mapped positions. The potential offered by GPS as a means of locating a sensor accurately and independently has already been considered as an important development in surveying and mapping of the late 1980s. The significance of this as a means of positioning satellite imagery is no less important than some of the other civilian applications planned for it. Nevertheless, at the time of writing, and probably for some decades to come, the user of MSS and similar imagery will have to depend heavily upon the availability of ground control. Moreover, whatever its merits, GPS can only be an aid in the location of an image if the aircraft or satellite is fixed by its means at the time of data collection. As long as we need to use older sources, obtained before GPS became available, we shall have to use the well-tried methods of locating position from ground control.

Although a satellite is a much more stable platform than any aircraft it is impossible to eliminate tilts entirely, so that corrections for tilt are required. As in air survey, it is difficult to make sufficiently accurate independent measurements of satellite attitude. For example, the gyro-

scopic tilt sensors on the early Landsat systems were accurate to only about $\pm 0°{\cdot}01$, which is insufficiently sensitive. The main difference from standard photogrammetric practice is that the methods of relative orientation used with conventional photogrammetry cannot be applied to Landsat imagery, first because these are not perspective photographs, secondly because there is no fore-and-aft overlap between adjacent scenes.

Although the ordinary methods of preprocessing are not particularly successful in recovering the absolute orientation of a scene, various forms of precision processing have been available at different times in the history of these satellites. For example, Landsat-4 and Landsat-5 TM images are processed in two forms of computer-compatible tape; the A-product tapes which are standard and the P-product tapes which are precision-processed. Sophisticated methods of pretreatment are also applied to SPOT HRV imagery, so that this is available at three different levels.

It should also be mentioned that there are local variations in the way that preprocessing is carried out. For example, the European Space Agency EARTHNET facility at Fucino in Italy, which is the principal receiving station for the Landsat imagery of Europe and North Africa, does not provide image data transformed to any projection, whereas all material processed in the USA by NASA is now fitted to the *Space Oblique Mercator* or SOM projection.

The geometry of the scanned image

The geometry of a satellite orbit has already been described in Chapter 16, pp. 377–379. Since the orbits of the early Landsat satellites were inclined at approximately 9° to the meridian, their heading was approximately 189°. In the following analysis we adopt the convention that the +x-direction is that of the path of the satellite and the +y-direction is that of the scanner sweep, perpendicular to this and transverse to the motion of the satellite. In a truly polar orbit this would mean that +x is directed to the south and +y points to the east.

The simplest illustration of the geometry of the single scan line is shown in Fig. 18.03, in which we may distinguish three different representations of the distance y from the sub-satellite point or nadir point for each scan line. This should be compared with Fig. 17.04, and the consideration of the effect of earth curvature upon the conventional instantaneous photograph. In this diagram, H is the satellite altitude above the earth's surface (which is 918 km for Landsat-1 through Landsat-3, and 705 km for Landsat-4 and Landsat-5); R is the earth's radius; and θ is the scan nadir angle. For the scan distance of the sensor, the maximum value of $\theta = 5°{\cdot}772$. Three distances are compared in Fig. 18.03:

- y_1 corresponds to distortion-free scanning.

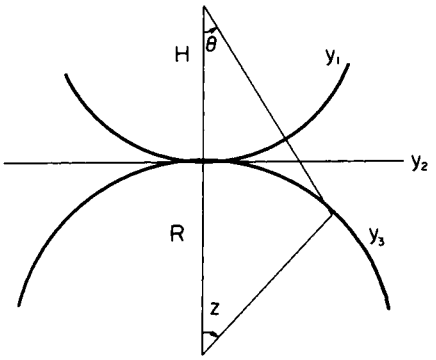


FIG. 18.03 Basic elements for the evaluation of the geometry along scan lines. (Source: Steiner and Kirby, 1976.)

- y_2 is the hypothetical ground distance for a flat earth and is, of course, the linear distance upon the plane of a projection which is tangential to the earth's surface at the sub-satellite point and the projection plane. This form of representation was considered in Chapter 17, pp. 372–376.
- y_3 is the real ground distance on a curved earth surface for any given value of θ. The angle z is the geocentric angle corresponding to y_3.

From Fig. 18.03 it follows that

$$y_1 = H \,.\, \theta = H \,.\, (at) \tag{18.01}$$

where a is the maximum value for θ at maximum scanning time t or $|\theta|_{max}/|t|_{max}$.

$$y_2 = H \,.\, \tan\theta = H \,.\, \tan(a \,.\, t) \approx H[at + (at)^3/3] \tag{18.02}$$

$$y_3 = R \,.\, z = R[\sin^{-1}(\Delta \,.\, \sin\theta) - \theta]$$

$$= R[(\Delta - 1)at + (\Delta^3 - \Delta)(at)^3/6] \tag{18.03}$$

$$= H[at + (\Delta^2 + \Delta)(at)^3/6] \tag{18.04}$$

where $\Delta = (H+R)/R$ as in equation (17.05). Obviously equations (18.02)–(18.04) correspond to those already determined for the planimetric influence of earth curvature discussed in Chapter 17. Because the motion of the rocking mirror system is not linear, it is necessary to modify the angle scanned, θ, to

$$\theta' = k_1 \sin(k_2 \,.\, t) \tag{18.05}$$

in these equations the constants $k_1 = 0{\cdot}29$, $k_2 = 21{\cdot}46$, $|\theta|_{max} = 0{\cdot}1005$ and $|t|_{max} = 0{\cdot}0165$ seconds, so that $a = 6{\cdot}0942$ and the maximum value

for $\theta' = 0{\cdot}10055$ radians. It follows that we may write

$$y_4 = H.\theta' = H.k_1 \sin(k_2.t) \approx H.k_1[k_2.t-(k_2.t)^3/6] \quad (18.06)$$

Consequently the individual imaging errors due to the non-linear functions, y_2, y_3 and y_4 are:

Panoramic distortion
This error is caused by scanning with a rotating mirror having constant angular velocity.

$$\Delta y_2 = y_2 - y_1 = (H.a^3/3)t^3 \quad (18.07)$$

Earth curvature

$$\Delta y_3 = y_3 - y_2 = (H.a^3/3)[(\Delta^2+\Delta)/2-1]t^3 \quad (18.08)$$

Non-linear sweep

which includes distortions in the optical system, non-linearity in the scanning mechanism and non-uniform sampling rates.

$$\Delta y_4 = y_4 - y_1 = H(k_1.k_2-a)t-(H.k_1.k_2^3/6)t^3 \quad (18.09)$$

From these equations we obtain the following combined errors:

Panoramic and earth curvature

$$\Delta y_5 = y_3 - y_1 = [H.a^3(\Delta^2+\Delta)/6]t^3 \quad (18.10)$$

All errors combined

$$\Delta y_6 = y_3+y_4-2y_1 = \Delta y_4+\Delta y_5 = H(k_1.k_2-a)t+H/6[a^3(\Delta^2+\Delta)-k_1.k_2^3]t^3 \quad (18.11)$$

Figure 18.04 illustrates the magnitude of these errors as referred to the time of scanning along one half of a single line between the nadir and the end of the scan line. It illustrates again the factor demonstrated for conventional aerial photography in Chapter 16; that the planimetric influence of earth curvature is relatively small. Figure 18.04 illustrates that the maximum displacement along each scan line occurs at the ends where it amounts to about 70 m, which is slightly smaller than the 79 m resolution of the single pixel of the early MSS scanners. It is, to the first approximation, the amount of deformation which we may expect to find in the single Landsat scene which arises from making the assumption that the earth is flat. It is, moreover, crucial to the arguments which follow that the other forms of error, including the skewing which is yet to be described, can all be eliminated by preprocessing before it is necessary to consider the influence of earth curvature and the need for a suitable projection.

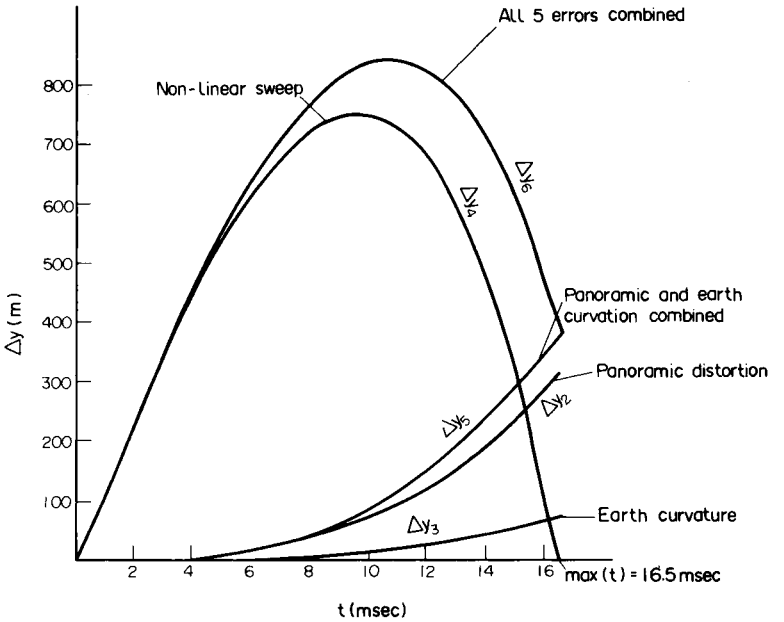


FIG. 18.04 Size of imaging errors along one half of a scan line as a function of scan time, t. (Source: Steiner and Kirby, 1977.)

The effect of earth rotation on MSS imagery

As the satellite travels southwards in its orbit it scans the earth from west to east across the narrow swath either side of its ground track. The rotation of the earth causes each successive mirror sweep to start a little further west than its predecessor. The overall geometric effect is to skew the image. Figure 18.05 shows the inherent distortion of an MSS scene. Depending on the latitude, the image is linearly expanded by the amount Δx in the down-track direction and linearly skewed by $H_e(\Delta y)$ in the cross-track direction. The quantity Δy varies with the cosine of the latitude and is therefore greatest at the equator. Non-linear distortion of the image is caused by the bend ΔH_e of the real sub-satellite track, affecting the gradually increasing lateral offset d of scan lines. The overall result of rotation is that the ground track is curved, as represented by the thick line in Fig. 18.06. The curvature varies sufficiently slowly for it to be possible to assume that it does not vary within the single scene. An important effect of earth curvature is the ground representation of the scan lines which are no longer parallel. Since these are normal to the ground track at any point, they must also converge or diverge as the ground track curves.

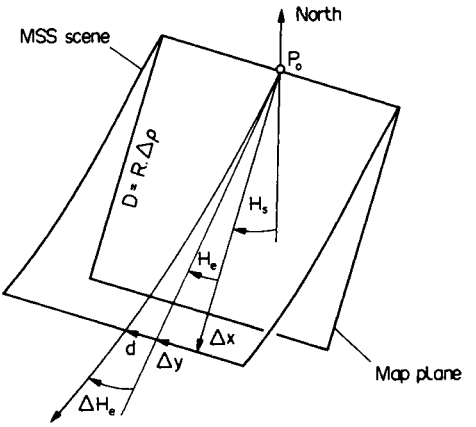


FIG. 18.05 The effect of earth rotation on MSS imagery. (Source: Kratky, 1974.)

The transformation of scanned images

In order to transform image data from the data tape into a map or GIS file it is necessary to convert the coordinates of the pixel in row r and column c into position on the earth's surface either as geographical coordinates (ϕ, λ), three-dimensional cartesian coordinates (X, Y, Z) or,

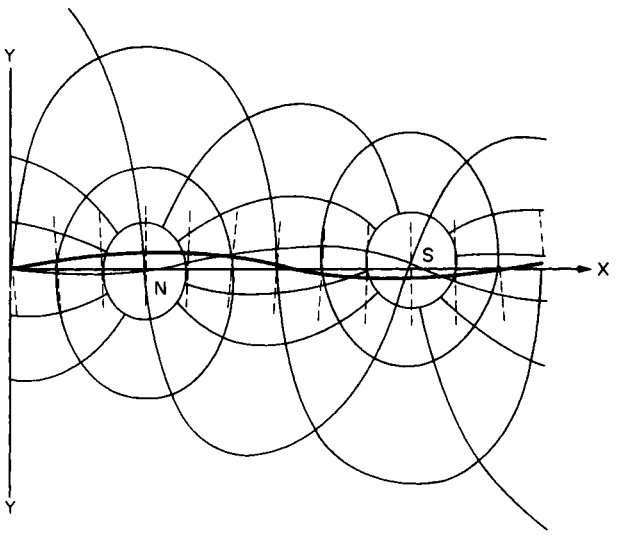


FIG. 18.06 The curved ground track of the Landsat satellite as it orbits the earth. The broken lines represent the direction of the lines scanned by the MSS, and which are always normal to the ground track. (Source: Snyder, 1981.)

finally, into the grid coordinates (x, y) or (E, N) for a plane map on a suitable projection. There are two approaches to making such transformations.

(1) The *analytical* or *attitude model* approach starts from a knowledge of the satellite ephemeris, that is to say the characteristics of the orbit of the satellite platform, combined with the earth's rotation and the sampling rates of the sensor. It is characteristically the photogrammetric approach to the problem, because the individual stages of the argument are examined in terms of these variables. Much of the early work on Landsat MSS was carried out this way, using photographs reproduced from the tapes so that the coordinates of image points could be measured in the traditional fashion by plotter, comparator or digitiser.

(2) The *numerical* approach is based upon the solution of two-dimensional transformation polynomial equations using interactive computer graphics to examine the data tapes. The coordinate transformations may be introduced through any of the variety of plane transformations using polynomials which, in the simplest form, are the conformal and affine transformations described in Chapter 2. The choice of the most appropriate method depends on the number of ground control points and their distribution within the image area. The numerical methods to be described are not unique to handling remotely sensed imagery, or, for that matter, to making maps. In the mapping sciences they also play a large role in the processing of GIS datafiles and in geodetic work, such as the change from one datum to another. In other branches of science and mathematics they are used for a variety of coordinate transformations. Chapter 19 is concerned with these applications, where the methods used are introduced in some detail. Here we offer only a brief summary of them, solely with reference to the treatment of scanned images.

Analytical transformation methods

We may express the transformation in the general form

$$(r, c) \rightarrow (\varphi, \lambda) \rightarrow (\mathrm{E}, \mathrm{N}) \tag{18.12}$$

or

$$(r, c) \rightarrow (\mathrm{X}, \mathrm{Y}, \mathrm{Z}) \rightarrow (\mathrm{E}, \mathrm{N}) \tag{18.13}$$

where (r, c) are the row and column pixel coordinates derived from the tapes containing the signals collected by the scanner, (φ, λ) are the geographical coordinates or (X, Y, Z) the three-dimensional cartesian coordinates defining terrestrial position and (E, N) are the map coordinates derived from them. An alternative is the process of registration where the (r, c) coordinates are transformed into an (x, y) system without

necessarily proceeding to the (E, N) grid coordinates, this procedure being repeated for multitemporal analysis so that each of several versions of the same scene can be compared on the common (x, y) framework.

Analysis of the attitude model function is exemplified by the study by Horn and Woodham (1979) who identify and quantify 12 variables. These are eventually combined as an affine transformation of the form

$$\left.\begin{aligned} E &= ax+by+c \\ N &= dx+ey+f \end{aligned}\right\} \tag{18.14}$$

where (E, N) are the grid coordinates and (x, y) are image coordinates derived from the (r, c) system. Many investigations were made in the early years of Landsat operations into the practicability and accuracy of mapping from MSS imagery; for example, Konecny (1976), Berrill and Clerici (1977), Steiner and Kirby (1976), Dowman and Mohamed (1981). It is interesting and significant that all of them reached virtually the same conclusion. For example, Berrill and Clerici described the comparison between the positions of checkpoints located by means of different transformation procedures and found that a simple linear affine transformation gave small random residual errors. They considered that no further improvements could be expected from the use of more complex interpolation methods, and they therefore concluded that limitations in planimetric accuracy obtained from Landsat images lay in the resolution of the imaging system itself.

Numerical transformation methods

An affine transformation such as (18.14) corrects first-order distortions such as angular deformation of the axes and the scale differences between along-track (x) and scan (y) directions which may be caused by earth rotation. However, the efficacy of simple affine transformation depends a great deal upon the extent of the geographic area which is to be mapped or registered to other images. Obviously no hard-and-fast rules can be applied to determine in advance how big an area should be treated in a particular fashion; only experience and experiment will serve as a guide. If the size of the area to be transformed appears to be too great to undertake as a single unit, then smaller blocks (or windows) must be created and transformed individually, or if this is less convenient than treating the scene as whole a higher-order polynomial, with a correspondingly greater number of control points must be used.

A second-order polynomial has the form

$$\left.\begin{aligned} N &= a_0+a_1c+a_2r+a_3c^2+a_4cr+a_5r^2 \\ E &= b_0+b_1c+b_2r+b_3c^2+b_4cr+b_5r^2 \end{aligned}\right\} \tag{18.15}$$

where N and E are the map coordinates, c and r are the image coordinates and a_1, b_1 are the coefficients being determined by the least-squares fit. This form of expression is used by the British National Remote Sensing Centre and has been described by Williams (1979) and Davison (1986).

The amount of data needed to determine the polynomial coefficients depends upon the *order* of the polynomial used in the rectification process.

A second-order polynomial equation having six terms such as a_0 through a_5 requires a minimum of six common points between the two systems (E, N) and (r, c) in order to obtain numerical values for all six coefficients. Polynomial transformation and interpolation methods are particularly well suited for handling remotely sensed data because of the importance of interactive computer graphics using a suitable digital image processing system. Interrogating digital data interactively also has important advantages over conventional interpretation techniques used with diapositives or paper prints which are viewed optically, simply because the multispectral scanners are far more sensitive than the human eye and the digital data recorded by them can reveal far more detail than a photographic print can show, or the human eye can detect. By employing many different image modifications which may be collectively described as *image enhancement*, it is possible to convert a rather indifferent monochrome picture with poor contrast into a clear multicoloured display. It follows that if the whole of the process, from identification of individual pixels through to mapping, can all be done using the same computer terminal rather than switching from digitiser or comparator to terminal and finally to a coordinatograph, there is economy both in the amount and variety of hardware required and in the number of different operator skills which are needed.

Choice of ground control on scanned imagery

The choice of ground control in conventional photogrammetric work was described in Chapter 17. Application of the same principles to the much smaller-scale images obtained from space must be tempered by the following limitations. First, the single scene covers a relatively large tract of country. Secondly, it is usual to select the ground control points from suitable detail shown on a small-scale topographical map and measure the (E, N) coordinates of these points on the map. On the other hand the regularity and precision with which orbits are repeated means that once a set of suitable ground control points have been located, they can be used time and time again. Therefore it is usual practice to create an archive of ground control points for future use. This method has been used in Britain by the National Remote Sensing Centre to select ground control points, and has been described by Davison (1986) and Benny (1983).

A factor which is commonly overlooked is that because the ground control points are points of map detail, their (E, N) coordinates have to be measured from a map. Moreover, the work is often done on fairly small-scale maps. Compared with the fairly elaborate procedures which are employed to make the images fit the map before selecting the (r, c) coordinates of a point, the assumption that the corresponding (E, N) coordinates have been measured on the map without error is probably the weakest link in the whole chain of operations.

The following factors may have affected the quality of the (E, N) coordinate measurements:

- the accuracy of how the feature to be measured has been represented on the map;
- the degree of cartographic generalisation which has been introduced in making a small-scale topographical map legible;
- the pointing accuracy of the reading microscope of a coordinatograph or the cursor of a digitiser;
- the nature of the source map used for measurement, whether this be a printed paper map, the original drawings or a copy of the map reproduced on a dimensionally stable base.

The different ways in which these factors affect linear measurements, including those made by coordinatograph and digitiser, have been treated in detail by Maling (1989). The first of these, notably planimetric accuracy, is unlikely to have much effect upon the coordinate measurements if a modern map is used. The other three are much less predictable, and are likely to give rise to errors in E and N which are far in excess of the limiting resolution of the scanned images.

Interpolation methods

The mapping functions which have been described can be used directly to transform the position of each image into a suitable map. However, this is seldom a practical solution for the enormous amount of computation required to evaluate these functions for millions of points. Because the amount of data to be transformed is so large, any technique which can be used to streamline processing is to be welcomed. Some economies may be achieved simply by using more economical ways of handling the data during computation. For example, Kratky (1975) has described analytical aspects of determining the unknown coefficients from data observed in regular two-dimensional grids, using a form of matrix manipulation due to Rauhala (1972). A brief introduction to this is given in Chapter 19.

Nevertheless a better solution is usually to reduce the amount of computation by creating an interpolation grid. The mapping functions can

be evaluated at the grid intersections and a bilinear interpolation technique can be used to map points within this control.

For example, Van Wie and Stein (1977) recommend the use of an interpolation grid comprising 20×20 lines or 400 equidistantly spaced interpolation points for each Landsat scene. This technique is much more economical in computing time because the full polynomial transformation need only be applied to the grid intersections forming the corners of a rectangle or quadrangle, and this is only a small fraction of the whole scene.

The other aspect of interpolation is that applied to resampling of scanned images, whether these be the products of remote sensing or raster scan digitising of an existing map. We have already seen that resampling techniques involve defining new pixel positions on the source image, or the map base, and filling these pixel positions with data chosen by one of the interpolation algorithms. There are three of these in common use:

- *Nearest-neighbour interpolation.* This preserves the radiometric quality of the original image but introduces localised geometric distortions which are discontinuities in the image. It is the most economical in computer time.
- *Bilinear interpolation.* This corrects the geometric distortions but acts as a low-pass filter introducing radiometric errors. It is the best compromise between expense and accuracy.
- *Cubic convolution.* This is considered to be geometrically the best method, but it is much slower than the others, needing almost twice as much CPU time as the nearest-neighbour interpolation.

Figure 18.07 illustrates a Landsat scene (broken lines) which is to be resampled in the process of fitting it to a map grid (full lines) and, in the process, establishing the digital number denoting the ground reflectivity of the pixel labelled α at the shaded cell on the superimposed grid.

Nearest-neighbour interpolation simply involves transfer of the digital number, α, to the whole pixel on the grid nearest to it.

Bilinear interpolation comprises transfer of the weighted overage of the digital number obtained for the four nearest pixels, being those cells labelled α and β in the image, to the shaded cell.

Cubic convolution comprises transfer of the weighted average of the digital number for the nearest 16 cells, these being the pixels labelled α, β and γ on the source image and transferring the result to the shaded cell. The relative speeds of computing these results are as follows: *Nearest neighbour*, 1; *Bilinear interpolation*, 10; *Cubic convolution*, 20. The reader is referred to descriptions of the methods in Williams (1979), Kratky (1981), Bernstein (1983), Burrough (1986), Richards (1986) and Mather (1987).

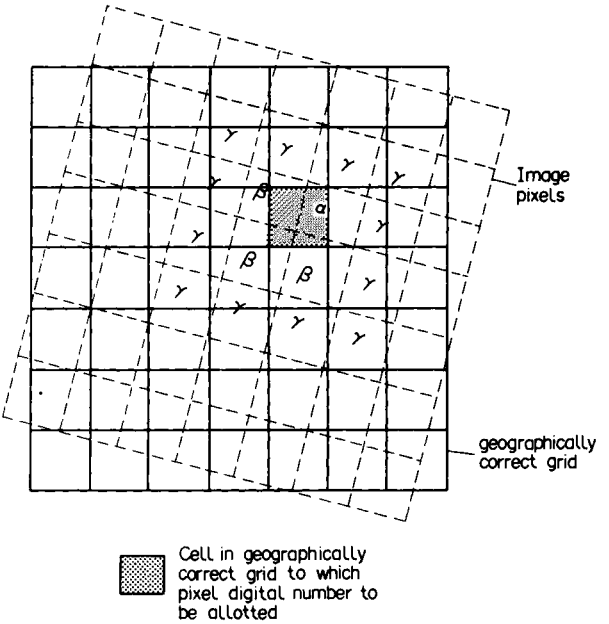


FIG. 18.07 Resampling procedure used to correct an image geometrically. (Source: Curran, 1985.)

Projections for Landsat MSS

The final stage in converting image to map is to consider the transformation of the corrected image coordinates into those on a suitable map projection. In Chapter 17 we demonstrated that the projection of a single vertical aerial photograph is a perspective azimuthal projection. However, a corrected scanned image does not possess the character of central perspective, so that the simple symmetry of the aerial photograph does not apply. We have seen that the single MSS frame is a raster composed of parallel scan lines each forming a row of pixels and this raster is an orthogonal net. When the various scanner and orbital perturbations have been corrected, distances along the line of scanning ought to be constant; so, too, is the separation between the scan lines. Because the orbital inclination is so close to a polar orbit, the effect of earth curvature upon image position may be expressed in rectangular spherical coordinates, as, for example, by Kratky (1974).

Projection distortion within the single Landsat MSS frame

If a single MSS frame is considered in isolation, the origin and ordinate may be considered to lie within that frame. Since the greatest deformation

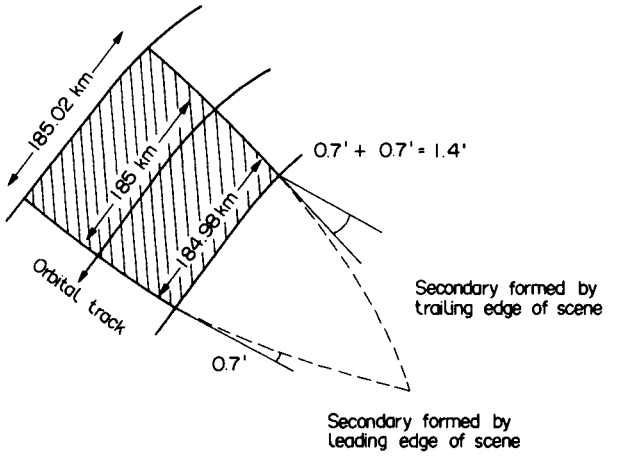


FIG. 18.08 The Landsat MSS scene considered in rectangular spherical coordinates. (Source: Kratky, 1974)

which can occur is in the direction parallel to the ordinate, and at the maximum distance to left or right from that ordinate, we imagine the ordinate to be the equivalent to the central meridian and located along the left-hand side of the frame. Then the greatest linear distortion will occur along the right-hand edge of the frame. Putting the orbital distance as 185 km, the maximum linear error attributable to earth curvature is approximately 19 m, or appreciably smaller than the side length of a pixel. There is also a very small amount of convergence between the secondaries corresponding to the first and last scan lines of the scene, but this amounts to only 1′·4, demonstrating that the influence of projection distortion is practically negligible, confirming the conclusion reached in equation (18.08) and illustrated in Fig. 18.04.

These values suggest that, provided the various geometric corrections to the scene have been applied in preprocessing, the residual errors owing to earth curvature are small enough to be neglected. This is, of course, equivalent to making the plane assumption in surveying, as described on pp. 321–323, but using a larger unit for the zero dimension corresponding to the resolution, and therefore the pixel size of the image. Because the zero dimension of an MSS image is one pixel side, the overall dimensions of the area within which the plane assumption is acceptable has increased in proportion. In practice, however, there is usually no need to consider the need for this application of the plane assumption in treating with Landsat scenes, for the choice of projection has already been made as a result of the methods of absolute orientation which have been described. Because the coordinates for ground control points are extracted by measurement from existing topographical maps, and because these maps

have already been prepared upon the projection used for the national survey, any map detail which is digitised with reference to the grid or graticule is automatically referred to that projection. It follows that any Landsat MSS scene which has been processed through the application of a polynomial expression such as (18.15) is already transformed into the same projection as that on which the ground control coordinates were measured.

If all the scenes used have been treated in the same way, preferably using the same control as well, there should be no difficulty in matching images for edge comparison or within the small overlap between scenes in order to obtain the necessary continuity through an area covered by more than one scene. However, we must assume that all the ground control points have been measured on the same projection, having the same datum. In the British Isles, for example, the ground control points located in England, Scotland and Wales are referred to the Transverse Mercator projection used by the Ordnance Survey, but those in Ireland are referred to a different version of the Transverse Mercator employed by the Ordnance Survey of Ireland, which has a different origin and is therefore a separate projection. Moreover, neither of these is compatible with the UTM. There are important differences between the OSGB36 or OSGB70(SN) datum used by the Ordnance Survey and ED50 in France and Belgium. Similar difficulties arise at the junction of scenes transformed to different control, for example, in the USA where some State coordinate systems are referred to Transverse Mercator projections which do not correspond to the UTM, or to the Lambert Conformal Conical projection. In all such examples it might be necessary to make a deliberate choice about which projection should be used for the scene or block of scenes as a whole, and how to fit this to the images. It means that the coordinates of some of the control points might have to be transformed to one of the other projections, but, in fact, there is usually no need for any correction. If the projections for the adjoining national surveys meet any of the requirements which have been established in Chapters 15 and 16, the discrepancies between coordinates in one system and those for the same point in the other seldom amount to more than 1 m on the ground. It follows that even if this is important in geodesy and surveying, in dealing with scanned imagery the differences are substantially smaller than the limiting resolution. Therefore they may be safely ignored for virtually all practical purposes.

A desirable projection for an entire strip of Landsat MSS imagery

As in conventional cartography, if the nature of the work extends over much larger areas, such as the production of a regional or national survey,

the importance of a special form of projection becomes apparent. Thus a large amount of effort has been put into deriving a suitable projection for use in preprocessing of Landsat or other imagery *as a continuous strip.*

It is necessary to emphasise this last statement, for the point is often overlooked that ever since Colvocoresses (1974) first investigated the problem, the object of choosing a suitable projection has been to serve as a base for an entire strip, such as mapping the single Landsat path from 81°N through 81°S as a single entity.

In effect there are three projections which have to be considered:

- the Transverse Mercator projection, particularly the UTM version of it, because this is so important as the base for topographical map series;
- the Oblique Mercator projection, because this may be assumed to fit the geometry of an inclined satellite orbit better than a transverse projection;
- the Space Oblique Mercator projection, which differs from the second by virtue of having curved lines of zero distortion corresponding to the true ground tracks of the satellite.

Colvocoresses (1974) argued that a suitable solution would have to employ the assumption that the earth approximates to the spheroidal shape, and also that any ellipticity of the satellite's orbit would have to be taken into consideration. The validity of both of these requirements was subsequently confirmed by Snyder (1978, 1981). In order to derive a suitably close relationship between the orbital path and a plane projection, Colvocoresses imagined a cylinder tangential to the earth as illustrated in Fig. 18.11, and made the initial proposal for a projection, called by him the *Space Oblique Mercator* or *SOM projection.*

A prototype version of SOM using the geometric analogy proposed by Colvocoresses (1974) was employed by NASA as a temporary measure until a more rigorous mathematical development had been achieved. This consisted basically of moving an obliquely tangent cylinder back and forth on the sphere so that the track around it which would normally be tangent shifted to follow the ground track. This is suitable near the equator but leads to errors of about 0·1% near the poles.

We have already seen that relating scanned imagery to the plane is complicated by the combined movements of the satellite and the earth's own rotation. The first and most important of these is that the ground track of the satellite can no longer be regarded as a great circle, but has the curved path illustrated in Fig. 18.09. Secondly, the effect of this curvature upon the scan lines is that these are no longer parallel; in some places they converge, elsewhere they diverge, as illustrated in Fig. 18.06. The individual scan lines, projected to the mapping plane, are mutually shifted and rotated, thus causing a variable distortion of the image

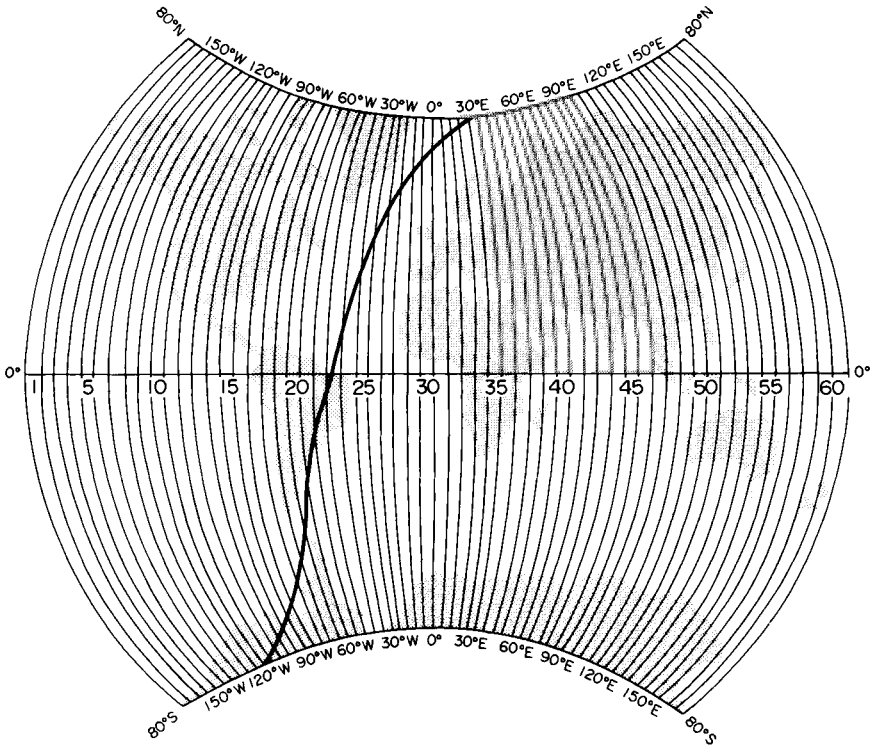


FIG. 18.09 The ground track of a single pass of Landsat superimposed upon the pattern of UTM zones.

geometry. In a rigorous analysis of cartographic errors one must examine how this already distorted pattern is further affected by other factors, the most important being the skewing effect which results from earth rotation during the period of scanning.

Transverse Mercator projection

We have already seen, in equations (18.01)–(18.11), how much deformation is present in a raw Landsat MSS image. Assuming that the various displacements, especially those caused by skewing of the image through earth rotation, have been effectively corrected in preprocessing, the remaining errors are random errors owing to the uncertainties of absolute orientation. If these, too, can be largely eliminated by use of ground control we are left with some small residual errors owing to the presentation of the ground track and scan lines on the plane of the projection.

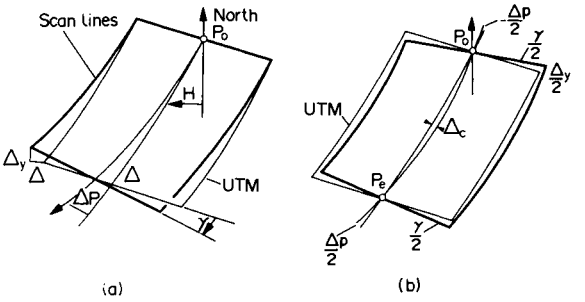


FIG. 18.10 UTM distortion in MSS images. (Source: Kratky, 1974.)

Figure 18.10(a) illustrates the theoretical changes in the geometry of an MSS image transferred in the course of the precision processing into the UTM plane. The primary angular bend of an MSS strip or of its part is additionally increased by the angle ΔP. This has been determined by Kratky to be less than 9 minutes. At the same time the direction of scanning gradually changes, thus causing the scan line convergence γ of the same magnitude, and the planimetric errors $\Delta < 240$ m, $\Delta_\gamma < 250$ m. As shown in Fig. 18.01(a) these errors represent absolute discrepancies accumulated within a frame, under the assumption that the position of the first scan line is correct. The influence of the errors is reduced because the MSS image is fitted with the UTM grid by using the full range of the scene. This leads to a situation illustrated in Fig. 18.10(b), where the points P_0 and P_e exhibit the best agreement, whereas the maximum residuals are distributed. The angular deviations are reduced down to the maximum of about 4′·5 and therefore the residual position errors are determined as $\Delta_c < 60$ m and $0{\cdot}5\ \Delta_\gamma < 125$ m. Whereas the error Δ_c may be tolerated, the displacement $0{\cdot}4\ \Delta_\gamma$ caused by the scan convergence is appreciable and should be taken into account in the precision processing of consecutive frames. The slight curvature of individual scan lines, which is theoretically present, is too small to be detected and may be neglected.

These ideas were included in the early imaging processing systems employed by NASA and, where processing to a projection was carried out before 1978, this was to the UTM. For example, Van Wie and Stein (1977) chose the UTM as the projection to which position should be referred in their design of the DIRS package for rectification of MSS imagery. Moreover, even in much later years, after abortive attempts to use an Oblique Mercator projection and the implementation of the Space Oblique Mercator projection, the UTM remains an important alternative solution. Thus in the production of P-quality CCT tapes by NASA from the MSS and TM imagery from Landsat-4 and Landsat-5, users still have

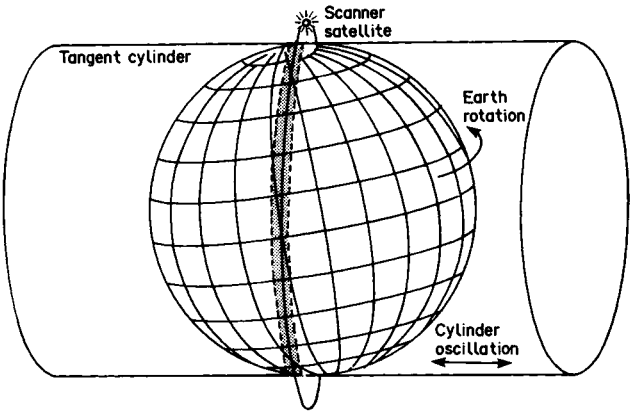


FIG. 18.11 Colvocoresses' concept of an oblique cylindrical projection surface which might take into account the effect of earth rotation, and therefore the curved ground track of a Landsat path. (Source: Colvocoresses, 1974.)

the option of using the UTM base with their own choice of ground control points.

Oblique Mercator projection

If the Transverse Mercator projection is considered to be unsuitable because of zone changes every 6° in longitude, an oblique aspect projection must be preferred, because the line of zero distortion may be oriented in a direction more closely corresponding to the ground track of the satellite. There is no difficulty in describing a suitable projection for the spherical earth. This has been done, for example, by Snyder (1981). However, greater difficulties occur in the creation of an oblique conformal cylindrical projection of the spheroid in which the principal scale is preserved continuously in the longer axis.

The form of the Oblique Mercator projection of the spheroid which has been used is that originally devised by Hotine (1946–7), which was used for mapping certain territories in the British Empire, for example Malaya and Sarawak, where the longer axis through the territory was aligned obliquely to the graticule. The projection is now known in North America as the *Hotine Oblique Mercator* or *HOM* projection. This projection proved unsuitable for two major reasons. First, as we have already seen, it makes use of double-projection from the spheroid to an auxiliary surface of an aposphere and thence to the plane. Although this is suitable for the representation of relatively short arcs, it cannot be extended for use through most of a hemisphere, far beyond the point where the spheroid and aposphere are tangential to one another. This is of little

consequence for mapping Malaya or Sarawak, but it is an objection to using it to map an entire Landsat path from 84°N to 84°S. Secondly, the curved groundtrack illustrated in Fig. 18.06 means that there cannot be good agreement between the line of zero distortion and the satellite groundtrack throughout.

Space Oblique Mercator projection

A rigorous algebraic study of the projection which results from Colvocoresses' visualisation of the problem was eventually done between 1975 and 1979 by Junkins and Turner (1978) and by Snyder (1978, 1981) working independently of one another.

The features of the SOM projection, which has been specially designed for continuous mapping of satellite imagery, are:

- It is a modified cylindrical projection having a curved line of zero distortion corresponding to the curved ground track illustrated in Fig. 18.07.
- It is intended only for use in a relatively narrow band along the groundtrack. This is because the projection is not rigorously conformal but the angular deformation cannot be detected anywhere within the zone occupied by the satellite images.
- The SOM requires minimal pixel resampling and consequently reduced computer processing time, both of which are very important considerations in the handling of the immense data load of TM.

We do not attempt to present the detailed algebraic arguments here, for these have been published in a variety of forms in Snyder (1978, 1981, 1982a, 1987a). He has produced a series of mapping equations based upon, first, the assumption that both orbit and earth are circular in section; secondly that the orbit is circular but the earth has spheroidal form and, third, that the orbit is elliptical and the earth's figure is spheroidal. The second assumption is valid for the later Landsat-4 and Landsat-5 orbits, but the third assumption is needed to make appropriate use of the elliptical orbit of the early Landsat satellites.

Although the SOM for the ellipsoid is not rigorously conformal, the error is negligible within the scanning range 50′ either side of the groundtrack. Scale in the direction of the groundtrack is correct for sphere or ellipsoid, while conformality is correct for the sphere and within 0·0005% of correct for the ellipsoid. At a distance of 1° from the ground track the particular scales vary between 1·000 154 and 1·000 150, corresponding to a percentage scale error of only 0·015% within any given Landsat path. The corresponding amount of maximum angular deformation, ω, is only 0·0006° in low latitudes ($\varphi < 25°$) falling to only 0·0001° in the vicinity of the geographical poles. For Landsat-4 and Landsat-5 the SOM pro-

jection has become standard for preprocessing, although the possibility exists for the user to request images fitted to the UTM.

The Space Oblique Mercator projection in practice

Welch and Usery (1984) and Welch *et al.* (1985) examined the geometric fidelity of the TM output from Landsat-4 and Landsat-5, and found the data significantly better geometrically than were obtained from the earlier Landsat missions. They found, not surprisingly, that a factor which may contribute to map errors is the difference between the UTM coordinates of a point on a map and its corresponding SOM coordinates, which suggests that all they have done is to apply a simple translation and rotation to the SOM values to obtain what they called pseudo-UTM coordinates. They argue that a more satisfactory solution is to convert the UTM coordinates into geographical coordinates using the inverse, or 'grid-to-geographicals' solution and then convert these back into another grid, this time having a local central meridian. It is, however, debatable whether these transformations really justify the amount of additional computing needed to make them. Although Welch claims that this discrepancy can now be detected under ideal conditions, it is insignificant compared to the total residual error, which is still primarily controlled by the spatial resolution of the sensor. Many of their results have been matched by other studies, e.g. Bryant *et al.* (1985). They found rather large discrepancies between the TM images and corresponding points referred to the UTM, but subsequent analysis showed that the projected image centre data computed from ephemeris information were in error. Indeed, they argue that the accuracy of a scene corrected to SOM is only as good as the orbit ephemeris data.

CHAPTER 19

Other methods of transformation

Mathematics alone cannot perform miracles
T. Vincenty, *Surveying and Mapping*, 1987

Introduction

This chapter extends the methods of coordinate transformation which have been introduced in Chapter 15 onwards to various applications in geodesy, surveying, topographic cartography and GIS manipulation. Special emphasis is placed upon the role played by such calculations in handling geographical information systems because these are rapidly becoming a major cartographic activity and show every sign of replacing many kinds of conventional map use in a decade or two.

Although a geographical information system may include all manner of positional information, we confine our attention to the task of transforming the detail from one map having an (x, y) grid to another map with an (X, Y) grid. We also make the assumption that the source is a conventional map. The great mass of positional data is stored in this fashion, and it will be a decade or two before the information systems grow sufficiently in scope and utility for them to have been created entirely from new mapping, which has never been subjected to conventional cartographic treatment. This means that the GIS will remain subject to all the limitations and disadvantages of the paper map irrespective of the sophisticated methods of collecting and handling of other data.

Required transformations

The various transformations which may be needed may be listed under the three following headings:

- change in medium;
- change in datum;
- change in projection.

Under the heading of *change in medium* there are, for example, those changes which may be caused by converting photographic or other remotely sensed images from the film or computer file into a conventional map or GIS file. These subjects have been sufficiently aired in Chapters 17 and 18 for there to be no need to consider them further.

Under the heading of *change of datum* we need to consider a variety of geodetic transformations which arise particularly in the execution of control surveys and fitting these to one another when one has been carried out independently of others. There are two principal components; a change in origin and a change of spheroid. There was a time when it was rare for a datum to be changed; the sheer labour of converting a whole network of control points by hand militated against this.

Before the development of digital computers, one of the largest tasks ever undertaken is also one of the least known, comprising the change by the USSR from geodetic control based upon the Bessel spheroid to the Krasovsky figure which, according to Zakatov (1962) was carried out at TsNIIGAiK (the Central Scientific Research Institute for Geodesy, Air Survey and Cartography) in 1942. Although the extent of control surveys throughout the Soviet Union was much less than it is today, one wonders how many people were involved in this task.

With the advent of digital computing it became a practical possibility to be more ambitious and combine independent and disparate national control surveys into a single unit. It is exemplified by the creation of the European Datum (ED50) carried out by the US Army Map Service in the immediate postwar years. More recent examples have arisen from the need to relate control surveys to the same datum as that used for fixing position by GPS. Consequently another major datum change has been that for control surveys in North America to what is now called the North American Datum of 1983, of NAD83. A description of this work has been given by Wade (1986).

Another important need for transformations of this kind arises in offshore surveying activities, for example in the North Sea, where places to be located with high accuracy often lie close to the median line which forms a maritime frontier between different countries which have land surveys based upon different origins and Figures of the Earth. Then it may be necessary to transform from one system to the other or convert both to a third system created specifically for the offshore work.

In addition to the obvious change from one projection to another, such as from Mercator's projection to the Azimuthal equal-area projection, *changes in projection* include:

- change in aspect, e.g. from normal to oblique aspect;
- change in scale-factor, which corresponds to the idea of the *modification* of a projection described in Chapter 11;

- change from one grid zone to another;
- change in the origin of similar grids;
- change in scale from one grid to another;
- change in orientation of grid axes.

It will be recalled that the last three changes are commonly combined in the grid-on-grid transformation introduced in Chapter 2.

The effects of the changes

A change in the position of a point resulting from a change in geodetic datum may, in some localities, amount to only a few millimetres on the earth's surface. Even in mapping from aerial photography, once the gross corrections to eliminate the effects of camera tilt and the influence of surface relief have been applied, the residual effect of earth curvature on image position is very small, and for purposes of mapping may safely be ignored. Similarly the projection distortion within the area covered by a typical frame imaged from a satellite is generally smaller than the resolution of the imaging system. In conventional cartography such small discrepancies in position are of little consequence because they are often far smaller than anything capable of being plotted on a map. In other words, we may use the zero dimension to fudge the final results long before we have to worry about the implications of such geodetic refinements. However, we have already seen that, in the use of the CORINE GIS files described in Chapter 12, there may not be such a clear distinction about what resolution may be significant. A most important degree of control is exercised by the nature of the source from which the file was originally prepared. We therefore reiterate the fundamental truth that no data can be made better than their sources.

Geometrical limitations of source maps

Since the sources of most positional data used in GIS files are conventional maps, it follows that the positional errors are those present in the source documents, to which must be added those which arise in digitising. The errors arise from a variety of different sources to which we have already referred in Chapter 18 where the problem was considered with respect to the accuracy of ground control points derived from topographical maps. These errors are so important in digital mapping and GIS manipulation that we make no apology here for returning to this subject, for it is desirable to emphasise that the problem is not confined to determining the (x, y) coordinates of ground control, but relates to every manner of converting mapped information into machine-readable form. First there are those errors which arise in making the map – what

is called quantitative map accuracy in Maling (1989), where the subject is treated in detail. The approach there is to study map accuracy with respect to cartometric measurements, and since vector digitising is a branch of cartometry, it has particular relevance in this context. The additional errors which may result from the digitising process result primarily from paper deformation of the source map, and operating errors in setting the cursor over it. Some of these errors may be reduced; for example, the effects of paper deformation can be wholly avoided by digitising the original map manuscript on a dimensionally stable plastic base. However, no matter how well this part of the work has been accomplished, there is always a limit below which errors on the source map cannot be reduced. This is, of course, the zero dimension which imposes the limit to legibility of a map. It follows that if we attempt to treat in an apparently rigorous fashion with a datafile created from conventional maps, we may indulge in a lot of inappropriate and time-consuming calculations which serve no really useful purpose. A typical example is that of using equations for the spheroid where those for the sphere would serve very well. Morrison (1989) has presented the conventional modern viewpoint of what can be done in computer mapping as follows:

> the difference between basing a map projection transformation on an ellipsoid as opposed to a sphere was in the past often not considered worth the added manual labour required to perform the more involved calculations. Moreover it was frequently regarded as nearly impossible to draft the improvements in the results of the calculations by manual means. With the computer doing the calculations and a high-resolution plotter drafting the results, the more accurate results from using an ellipsoid can now be achieved as easily as the less accurate results based on as sphere. Therefore, the criteria used by the cartographer in taking a decision of the particular procedures to be used in making a map are also changed.*

There is no problem in writing and executing programs which apply transformations with geodetic precision, and such practices are appropriate to field surveys and simulated maps. However, they are extravagantly time-consuming to apply to handling GIS layers derived from paper maps, for which a spherical solution is still appropriate. Snyder (1985, 1987b), Shmutter (1981) and Doytsher and Shmutter (1981) have presented formulae for transforming existing map data to and from various projections. All of these solutions have been derived for a spherical earth, and they are considered to be sufficiently accurate for use with such sources.

*I have had to take a liberty with this quotation because Morrison twice uses the word 'spheroid' where clearly he means 'sphere', and we have already established that the words 'ellipsoid' and 'spheroid' are synonymous. However, this slip of the pen does not detract from the implication of the statement that better results should come from using the spheroidal assumption.

Grid cells and GIS frameworks

In handling GIS files of continental or even world dimensions a key activity is to make satisfactory comparison between different layers. This has to be executed quickly and efficiently, notwithstanding the fact that the files may hold data concerning millions of other places. The data structure of such files, and the means of accessing them, it not relevent to the present book. The reader is referred to standard works on GIS, for example Burrough (1986), Jackson and Mason (1986) and several papers presented at the AUTOCARTO LONDON conference, which have been published in Blakemore (1986) on the subject. It will suffice to state that, in the late 1980s there seemed to be consensus favouring organisation of the data into *quadtrees*.

Our preoccupation here is with the methods adopted for coordinate referencing, and the properties of some of these for ease of access and storage of information. For GIS purposes position is usually recorded on the *spherical surface* in one of three ways:

- by geographical coordinates,
- by a system of *grid cell* reference,
- by using a modification of geographical coordinates to isometric (conformal) or authalic (equal-area) coordinates.

On the *plane* a variety of grid or projection coordinates, or combinations of these may be employed.

In the present context we are concerned with continental or world-wide geographical information systems and, it should be emphasised, not land information systems which have only local applications for which a national grid system or the UTM will suffice. The principal requirements for a global coordinate system may be listed as follows:

- A hierarchical data structure is required in order to store data at different levels of resolution, and a regular hierarchy, using equal numbers of subdivisions, tends to be more efficient than an irregular hierarchy. This is examined in some detail by Jackson and Mason (1986).
- It is obviously necessary to be able to transform positional data from the source document to the GIS, but less obvious is the need to transform from the GIS back again to the source map. However, this is necessary, for example, when we convert the coordinates digitised from a map into some resident system in the GIS and then reverse the procedure in order to recover and plot the original map showing some additional information extracted from other files within the GIS.
- The system should have a clear and simple relation to geographical coordinates, for these still provide a reference system which is not

truncated by the artificial boundaries of a grid, or the form of *grid cell* reference system to be found in most atlases where the numbering system is unique to only one or two pages in that particular atlas.

- Since the primary requirement of a continental or world GIS is to create an inventory of geographical phenomena, we argue that preservation of area is of greater importance than absence of angular deformation, and that therefore the world should be partitioned into chunks of equal size.

The subject of geographical coordinates need not detain us here, for we have been concerned with their presentation or use on almost every page of this book. However, the concept of the grid cell needs a few additional words of explanation. This is really an alternative name for the spherical quadrilateral or quadrangle, used in earlier chapters. We introduce the method by means of a simple example, which has been described by Cocks *et al.* (1988) in the development of AIS, the geographical information system for the whole of Australia. This system uses grid cells of dimensions $1^\circ \times 1^\circ$ and $\frac{1}{2}^\circ \times \frac{1}{2}^\circ$, as illustrated in Fig. 19.01. Statistical and other data obtained from other sources first have to be referred to these quadrangles. However, it is important to bear in mind the low resolution

FIG. 19.01 The grid cell system used for referring AIS census data in Australia. (Source: Cocks *et al.*, 1988.)

of the system if subdivision is limited only to $\frac{1}{2}°$ cells. Imagine the validity of a relief map of the continent if the only available data were one spot height in each quadrangle. Tobler (1988, 1989) has investigated the correspondence between size of grid cell, limiting resolution of a system and the map scale to which these would apply. He concludes that for a $1° \times 1°$ cell the limit of resolution (which Tobler puts at 0·5 mm) provides an equivalent map scale of 1/17 800 000. He therefore concludes that if the minimum detectable size is twice the resolution, at that scale a country the size of Switzerland can hardly be detected.

Even if the resolution is improved by employing much smaller grid cells, the principal disadvantage of making a simple subdivision of the graticule is that the grid cell is, or course, regular neither in shape nor in size. Because each cell is formed from two convergent meridians and two parallels the shape of the figure is not a rectangle. Because convergence varies with latitude, so does the shape and the area of the figure. In a continent the size of Australia the area of the $1° \times 1°$ grid cell varies from about 12 000 km^2 in the north (latitude 10°S) to 9000 km^2 in latitude 43°S in southern Tasmania; a difference of about 3000 km^2 or nearly 30%.

There are other more complicated methods of subdividing the spherical surface. For example Tobler and Chen (1986) refer to work carried out in 1981 for the National Telecommunications and Information Administration of the US Department of Commerce, which involved breaking down the spherical surface into: 36 *zones*, each successively subdivided into 3060 *regions*, 15 *districts*, 64 *blocks* and 22 801 *points*. A 'point' in this system is a small figure measuring three arc-seconds in both latitude or longitude, approximating at the equator to a square of sides 93 m. There are $2{\cdot}4 \times 10^{12}$ of these to be input, stored and extracted for processing if the whole world is treated with similar detail in one gigantic datafile. The 93 m point is approximately 2–3 times the area of the unit pixel employed with the Landsat TM and SPOT imagery. The same disadvantages of variable shape and size still apply to the 93 m point mapping unit, but because we are so accustomed to think in terms of the zero dimension we regard such a small subdivision of the spherical surface as being a uniform square. Obviously, however, the larger units of the hierarchy are affected by the characteristic irregularity of the spherical quadrangle. Therefore somewhat between the levels of magnitude of districts, blocks and points, the idea that a small cell has uniform dimensions becomes a demonstrably untenable assumption.

The present author has argued in Maguire *et al.* (1991) that some kind of resident projection system or *GIS Framework* is the desirable alternative. A conventional approach has been adopted by the author in choosing and designing a resident projection to be used as the GIS framework for the CORINE environmental GIS of the European Community, details of which have already been given in Chapter 12. For such

an area, which is small by the standards of most of the other continents, a number of equal-area projections might be used without creating any excessive distortions within the area of study. Thus it was shown that the preference for either the Azimuthal equal-area projection or Albers' projection is largely academic once the most suitable origin and aspect have been selected. The corresponding choice for a map of the whole of the Old World, for a hemisphere, or, most difficult of all, for the whole world, is less easy and a conventional map projection is unacceptable when part of the data relates to places which are not far removed from singular points. Consequently less conventional projections may be needed. Tobler and Chen (1986) have indicated one possible solution is to use polysuperficial or multi-faceted maps comprising a pattern of recentred projections based upon the platonic solids, namely the projections of the world represented on the faces of a cube, octahedron, dodecahedron and icosahedron – and in which each facet is separately gridded. A particularly useful example may be the *Square equal-area map of the world*, devised by Gringorten (1972). There is no reason why a pattern of such separate projections should not serve as the basis of a world GIS, apart from the likelihood that it would be difficult to handle data lying at the common boundaries of the facets and ensure that data can be compared across gaps which do not occur on the earth.

Using regular gores rather than facets, Mark and Lauzon (1985) proposed a global scheme based upon the gores formed by the UTM map zones illustrated in Fig. 16.07. The specific proposal was to cut each UTM into subzones and then into patches, each subdivided into patterns of 256×256 pixels of 30 m side length. The method has the advantage that the projection system is already the basis of the most important projection used for topographical mapping. However, it shares with the other polysuperficial subdivisions of the world the disadvantage that the boundaries between UTM zones (such as the Greenwich Meridian) become major discontinuities which may be difficult to bridge.

The Authalic Grid

Tobler and Chen (1986) have further considered solutions making use of alternative coordinate systems upon the spherical surface. We have already made use of isometric coordinates in order to create conformal maps. In Chapter 16, however, the purpose was to map the spheroid conformally upon a sphere. Similarly a system of authalic coordinates may be used to create equal-area cells. In this context the purpose is to provide a graticule which subdivides the whole of the curved surface into cells of equal size. The simplest of these equal-area coordinates modifies only the spacing of the parallels by introducing an authalic latitude

$$\varphi' = \sin \varphi \qquad (19.01)$$

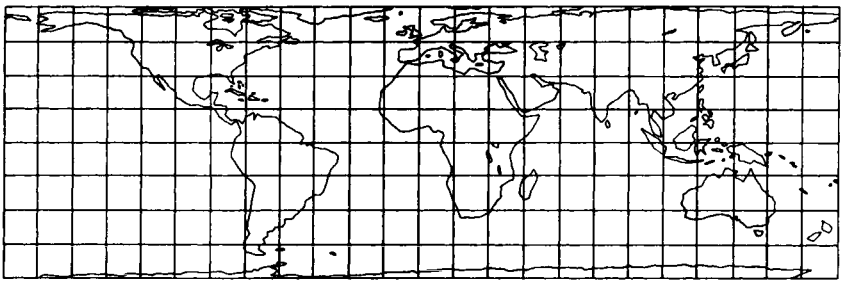


FIG. 19.02 The Cylindrical equal-area projection with an Authalic Grid. Compare this with Fig. 6.04. (Source: Tobler and Chem, 1986.)

This is equivalent to the use of the Cylindrical equal-area projection Fig. 6.04, p. 112, on which a system of equal-sized squares can be drawn, and is illustrated in Fig. 19.02. The difference between the two figures is the location of the parallels. In Fig. 19.02 they are a uniform distance apart on the sphere, and in the other case they are spaced to obtain quadrilaterals of equal area. These quadrilaterals become the square nodes of a quadtree.

The transformation methods

The transformation from one map projection to another is essentially the coordinate conversion of a point on one plane to the corresponding point on another plane. Then the basic equations for transforming the coordinates of one point to those of another can be written as

$$x = f_1(\varphi, \lambda); \qquad y = f_2(\varphi, \lambda) \tag{19.02}$$

$$X = F_3(\varphi, \lambda), \qquad Y = F_4(\varphi, \lambda) \tag{19.03}$$

There are two basic methods of undertaking the work which we shall refer to as: analytical or indirect transformation, and direct or numerical transformation. Much of the cartographic theory which follows seems to have been investigated by Kavraisky before digital computing had even become a practical possibility. The computer applications were described by Pavlov (1967) and this work was closely followed in China by Wu and Yang (1981). Their contribution was subsequently greatly extended by Snyder (1985).

Analytical transformation

This is the most straightforward solution to the problem, and we have already indicated that it is the obvious way of maintaining a uniform

collection of data from different sources, just as a world gazetteer or atlas index is likely to use geographical coordinates as the means of referring to position. The object of the transformation is to convert back from the coordinate positions digitised on the source map into their geographical coordinates. These, in turn, are used to determine the projection coordinates for the new map. It is called the analytical method because it employs the kind of equations derived in Chapter 10. The conversion from geographical coordinates into plane coordinates which we have regarded as the normal practice in map projections we may now regard as the *forward* equations. Those needed to determine geographical coordinates from a given map are the *inverse* equations. If the functional expressions for the original projection are those of equations (19.02) and (19.03), the simplest transformation model is:

$$\underset{\longleftarrow\text{INVERSE SOLUTION}\longrightarrow}{(x, y)} \rightarrow \underset{\longleftarrow\text{FORWARD SOLUTION}\longrightarrow}{(\varphi, \lambda)} \rightarrow (X, Y) \tag{19.04}$$

As an example of the relationship between the forward and inverse coordinate expressions, we consider the equations for Mercator's projection. For the forward solution, equation (10.64) applies. In order to express (φ, λ) in terms of (x, y), which is the inverse solution, we write, for the spherical assumption

$$\begin{aligned} \varphi &= \pi/2 - 2\arctan(\varepsilon^{-y/R}) \\ \lambda &= x/R + \lambda_0 \end{aligned} \tag{19.05}$$

where ε is the base of natural logarithms ($= 2{\cdot}1782818\ldots$). It is here written as the Greek epsilon to avoid confusion with the eccentricity of the spheroid, e, in the next equation. The term λ_0 represents the datum of longitude measurement.

For inverse solution of Mercator's projection of the spheroid we have to modify equation (10.82). The result is an equation of the form

$$\varphi = \pi/2 - 2\arctan\{t[(1 - e \cdot \sin\varphi)/(1 + e \cdot \sin\varphi)]^{e/2}\} \tag{19.06}$$

and $t = \varepsilon^{-y/a}$ which requires an iterative solution.

For the first trial we put

$$\varphi = \pi/2 - 2\arctan t \tag{19.07}$$

The result is inserted as φ in the right-hand side of (19.06) to calculate a new value for φ on the right-hand side. The process is repeated until the results have converged, and the user considers the difference between two successive determination of φ to be insignificant.

Longitude is obtained from a simple modification for the λ expression in (19.05), namely

$$\lambda = x/a + \lambda_0 \tag{19.08}$$

The reader will find the forward and inverse equations for most of the commonly used map projections listed in Snyder (1987a), together with worked examples of the solutions for both the sphere and spheroid.

Additional complications

The transformation model (19.04) is the simplest example. Some of the following complications may arise even in quite simple examples. The first stems from the fact that most digitising is done in cartesian coordinates, but we may need to deal with conical or azimuthal map projections which would be better derived in polar coordinates. Then it is necessary to change the plane rectangular coordinates (x, y) of the digitiser into plane polar coordinates (ρ, δ) determining the geographical coordinates. This is quite simply done through equations (2.01) and (2.02) and

$$\underset{\longleftarrow\text{INVERSE SOLUTION}\longrightarrow}{(x, y) \rightarrow (\rho, \delta)} \underset{\longleftarrow\text{FORWARD SOLUTION}\longrightarrow}{\rightarrow (\varphi, \lambda) \rightarrow (X, Y)} \qquad (19.09)$$

In changing the aspect of a projection an additional step must be taken after the geographical coordinates have been obtained. From the description of the method in Chapter 9, these have to be transformed into spherical polar, or bearing and distance coordinates, (z, α) as we have called them, using equations (9.04) and (9.05). Finally these have to be converted into plane (X, Y) coordinates using the projection equations expressed in terms of z and α, as in (9.10).

$$\underset{\longleftarrow\text{INVERSE SOLUTION}\longrightarrow}{(x, y) \rightarrow (\rho, \delta)} \underset{\longleftarrow\text{CHANGE IN ASPECT}\longrightarrow}{\rightarrow (\varphi, \lambda) \rightarrow (z, \alpha)} \underset{\longleftarrow\text{FORWARD SOLUTION}\longrightarrow}{\rightarrow (X, Y)} \qquad (19.10)$$

We have also seen in Chapter 9 that an important alternative to the (z, α) method of changing aspect is through the application of rotations to the three-dimensional cartesian coordinates of a point. Then we need to make the transformations

$$\underset{\longleftarrow\text{INVERSE SOLUTION}\longrightarrow}{(x, y) \rightarrow (\varphi, \lambda)} \underset{\longleftarrow\text{CHANGE IN ASPECT}\longrightarrow}{\rightarrow (X, Y, Z) \rightarrow (X^*, Y^*, Z^*)} \underset{\longleftarrow\text{FORWARD SOLUTION}\longrightarrow}{\rightarrow (\varphi', \lambda') \rightarrow (X, Y)} \qquad (19.11)$$

where we have put the three-dimensional matrix transformation described on pp. 185–194, in italics to avoid confusion with the use in these equations of (X, Y) as the master grid coordinates of the transformed points.

The advantages and disadvantages of the analytical method

There was a time when the indirect method was not only the most obvious but also virtually the only way of tackling the problem, particularly when

working on the spheroid. For most survey applications it was essential to employ projection tables similar to those mentioned in Chapter 16. Since these tables were based upon the geographicals-to-grid and the grid-to-geographicals solutions this was virtually the only way of computing changes in projection, changes from one grid zone to another, and changes in datum. For example, Zakatov (1962) provides a comprehensive account of the methods of changing zone in SURS which is wholly based upon making this transformation. It is still used for many survey applications, e.g. Field (1980), who has described the transformation of survey control from the Nigerian version of the Transverse Mercator projection (NTM) into UTM, using the analytical method.

Apart from the fact that the method is rigorous and is independent of the size of the area to be mapped, it has, nevertheless, three major disadvantages.

- information about the projection of the source map may be incomplete or even non-existent;
- the method can be inconveniently slow because so many individual coordinate conversions may have to be applied to each point;
- at large scales it may be wholly irrelevant to use geographical coordinates.

Snyder (1985, 1987c) has argued that the labelling of projections on existing maps leaves much to be desired, and that even when correctly named, important information such as the positions of the standard parallels in a conical projection or the central meridian of the particular version of the Transverse Mercator projection have not been stated. We might be expected to assume, of course, that the projection of the final map is known, so that the final transformation into (X, Y) coordinates can be correctly specified. However, Snyder suggests that even this is not necessarily so. In order to overcome some of the difficulties which arise from incorrect description, he has written a program which attempts recognition of a projection in use which based upon the digitised coordinates of nine points (on three parallels and three meridians) of the map, but even this can only distinguish between fairly simple examples. It seems to be of relatively little value in making specific identification of the projection used for topographical maps and aeronautical charts, which is the field of cartography where many of these transformation problems arise. In extreme cases, absence of information about the projection of the parent map may prevent use of the analytical method.

The analytical method may be inconveniently slow because so many different and separate coordinate conversions may have to be applied to each point on the map. For example, in (19.09) eight, and in (19.10) 10 separate transformations are needed to convert from (x, y) to (X, Y). Nowadays speed of computation ought not to be a problem. Perhaps we

should agree with Vincenty (1985), that modern high-speed computers with virtual memories should reduce these considerations to insignificance, and that:

> To expect a cost reduction from the use of faster transformation formulae is like offering the contents of a child's piggy bank to help to reduce the national budget deficit.

But the object of most digital mapping is not to compute the transformation of just one point, or even a graticule comprising several hundred points. In order to dispense with proportional dividers in compilation it is necessary to apply the same transformation to all the map detail, and this may well involve repetition of the entire procedure hundreds of thousands or even millions of times. Then the analogy is that several million piggy banks may, indeed, have an effect upon the budget deficit. Excessive processing time may be reduced by simplifying the equations and using a spherical solution, as already suggested, but this is only a partial palliative. The real solution must come from using one of the other methods or even using the transformation equations for only a sparse network of 'control' points, using interpolation procedures to locate the detail within the network formed by these. This subject was briefly considered in Chapter 18.

The final objection to the use of the analytical method of transformation is that, at larger scales, we are working with what are effectively two plane fields; between the photograph and the map, between one map and another. The majority of large- and medium-scale maps are prepared wholly upon grid systems such as the National Grid of Britain, the UTM or a State coordinate system. Although each of these can ultimately be related to the system of geographical coordinates, the user is often unaware of, or indifferent to, this fact. Therefore the use of geographical coordinates as the common medium for all data may be irrelevant so that conversion to and from geographical coordinates becomes a shocking waste of time.

Direct transformation

The method does not require transformation into geographical coordinates and back to the grid of the new map, but is based upon the relation between the rectangular coordinates of the same points on the two grids. This subject was first introduced in Chapter 2 in the simplest form, the very name grid-on-grid indicating its purpose. Therefore the model of the normal transformation is

$$(x, y) \rightarrow (X, Y) \tag{19.12}$$

The methods described here play a leading role in numerical analysis and have many other applications which have nothing to do with making maps.

We have already considered two simple methods of grid-on-grid transformation which we have already had occasion to use for several purposes:

- the linear conformal, similarity or Helmert transformation, described originally in Chapter 2 and expressed by equations (2.09);
- the affine transformation, expressed by equations (2.10).

These are *linear* or *first-order polynomials*, which are adequate for making many kinds of simple transformation. We shall find that they occur as the lowest-order terms in much more elaborate polynomial transformations which have to be employed if there are more complicated functional relationships between corresponding points.

The types of transformation which need higher-order polynomials are generally those requiring a more accurate result, or those in which the relationship between the two surfaces is particularly complicated. For example the transformation of the geodetic datum for the North American continent from NAD 27 to NAD 83 needs high accuracy results for it to have any practical utility. On the other hand the transformation from scanned images generated by the Landsat TM and SPOT systems, described in Chapter 18, needs the use of higher-order polynomials, not so much for reasons of high accuracy but to remove some of the smaller geometrical distortions to relate the images to a conventional map or GIS file.

The direct numerical transformation is also needed if the analytical equations for the original projection are unknown, or are uncertain, and it may be impossible to calculate the $(x, y) \rightarrow (\varphi, \lambda)$ relationship.

A first-order transformation may suffice for some simple examples, but those involving manipulation of the Transverse Mercator projection need a second- or third-order polynomial.

The derivation of a complex polynomial for conformal mapping

In order to indicate how a general polynomial for conformal mapping may be derived, we consider that derived through the medium of complex algebra.

We start from the functional relationship introduced in equation (16.19) and for the two grids we may write

$$x + iy = f_1(q + i\lambda) \tag{19.13}$$

$$X + iY = f_2(q + i\lambda) \tag{19.14}$$

where, as before, q is the isometric latitude and i is the complex number. Then we eliminate $(q + i\lambda)$ and obtain a direct transformation of the form

$$X + iY = F(x + iy) \tag{19.15}$$

Expressing this in higher-order terms, we may put

$$(X+iY) = (a_0+ib_0)+(a_1+ib_1)(x+iy)+(a_2+ib_2)(x+iy)^2 \ldots \quad (19.16)$$

or in a more generalised form

$$(X+iY) = \Sigma(a_k+ib_k)(x+iy)^k \quad (19.17)$$

Putting k = 1, the two terms expand to

$$(X+iY) = a_0+ib_0+a_1x-b_1y+ib_1x+ia_1y \quad (19.18)$$

and equating real and imaginary parts

$$X = a_0+a_1x-b_1y \quad (19.19)$$

$$Y = b_0+b_1x+a_1y \quad (19.20)$$

which are the now familiar expressions for first-order conformal transformation (2.09). Putting k = 2, for a second-order polynomial, the equations become

$$X = a_0+a_1x-b_1y+a_2(x^2-y^2)-b_2xy \quad (19.21)$$

$$Y = b_0+b_1x+a_1y+b_2(x^2-y^2)+a_2xy \quad (19.22)$$

Equations of this form have been used by Lucas (1977) to transform from a local control network to UTM, by Olliver (1981) to compute changes in UTM zone, and by Graff (1988) as one way of transforming from NAD 27 into NAD 83.

Third-order polynomial expressions relating grid to geographical coordinates may be written in the form:

$$X = a_{00}+a_{10}\lambda+a_{01}\varphi+a_{20}\lambda^2+a_{11}\lambda\varphi+a_{02}\varphi^2 + a_{30}\lambda^3+a_{21}\lambda^2\varphi+a_{12}\lambda\varphi^2+a_{03}\varphi^3 \quad (19.23)$$

$$Y = b_{00}+b_{10}\lambda+b_{01}\varphi+b_{20}\lambda^2+b_{11}\lambda\varphi+b_{02}\varphi^2 + b_{30}\lambda^3+b_{21}\lambda^2\varphi+b_{12}\lambda\varphi^2+b_{03}\varphi^3 \quad (19.24)$$

Those used to transform from grid to grid are:

$$X = c_{00}+c_{10}x+c_{01}y+c_{20}x^2+c_{11}xy+c_{02}y^2 + c_{30}x^3+c_{21}x^2y+c_{12}xy^2+c_{03}y^3 \quad (19.25)$$

$$Y = d_{00}+d_{10}x+d_{01}y+d_{20}x^2+d_{11}xy+d_{02}y^2 + d_{30}x^3+d_{21}x^2y+d_{12}xy^2+d_{03}y^3 \quad (19.26)$$

These and higher-order polynomials (up to the sixth degree) have been tried by Vincenty (1987) for the NAD transformation, and we have seen that the third-degree polynomial is often used for geometrical correction of remotely sensed imagery.

In pre-computer days polynomial expressions were usually left in this form because it was generally easier to compute each term individually. However, in view of what has already been said about economy in the design of equations, a nested form of each equation may be obtained from a little algebraic rearrangement. For example, the expression for x in (19.23) may also be written as:

$$X = a_{00}+\varphi(a_{01}+a_{02}\varphi)+\lambda(a_{10}+\varphi(a_{11}+a_{12}\varphi)) + \lambda^2(a_{20}+a_{21}\varphi+a_{30}\lambda)\ldots \quad (19.27)$$

This example is particularly instructive. Snyder (1985) has reported that the savings which result from using (19.27) rather than (19.23) are between 20% and 30% in the solution of a fifth-order polynomial.

Determination of the polynomial coefficients

The number of common points for which both (x, y) and (X, Y) or (φ, λ) are known, and which are needed to establish the coefficients for the polynomial, varies according to the order or degree of the polynomial. Thus, first-, second-, third-, fourth- and fifth-degree polynomials require a minimum of 3, 6, 10, 15 and 21 corresponding points respectively. If more control points are available than are needed for a polynomial of particular order, then the coefficients may be determined by the following least-squares method.

This ($m \times n$) matrix solution is applicable for any number of coefficients, n, and common points, m, but a practical limit is usually created by the capacity of the computer. It is well known in numerical analysis that although a polynomial may be extended to include higher-powered terms in φ^4, λ^4, φ^5, λ^5, ..., etc. the labour of determining the coefficients will hardly justify the extra computing time. Snyder (1985) provides an example which shows that increasing the degree of the polynomial from third-order to fourth-order barely justifies the greater accuracy obtained for any purpose other than geodetic work.

In equations (19.28) and (19.29) the individual coefficients form the column matrix on the left-hand side and the control, or common point coordinates are the column matrix on the right-hand side.

$$\begin{bmatrix} a_{00} \\ a_{01} \\ \ldots \\ \ldots \\ a_m \end{bmatrix} = \mathbf{D}. \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ \ldots \\ x_m \end{bmatrix} \quad (19.28)$$

$$\begin{bmatrix} b_{00} \\ b_{01} \\ \dots \\ \dots \\ b_m \end{bmatrix} = \mathbf{D} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ \dots \\ y_m \end{bmatrix} \tag{19.29}$$

The matrix **D** is calculated from

$$\mathbf{D} = [\mathbf{A}^T \cdot \mathbf{A}]^{-1} \cdot \mathbf{A}^T \tag{19.30}$$

where the $(m \times n)$ matrix **A** is formed from the geographical (or grid) coordinates of the corresponding points. Thus for the third degree polynomial requiring ten terms per line, or $n = 10$

$$\mathbf{A} = \begin{bmatrix} 1 & \lambda_1 & \varphi_1 & \lambda_1^2 & \lambda_1\varphi_1 & \varphi_1^2 & \lambda_1^3 & \lambda_1^2\varphi_1 & \lambda_1\varphi_1^2 & \varphi_1^3 \\ 1 & \lambda_2 & \varphi_2 & \lambda_2^2 & \lambda_2\varphi_2 & \varphi_2^2 & \lambda_2^3 & \lambda_2^2\varphi_2 & \lambda_2\varphi_2^2 & \varphi_2^2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \lambda_m & \varphi_m & \lambda_m^2 & \lambda_m\varphi_m & \varphi_2^2 & \lambda_m^3 & \lambda_2^2\varphi_m & \lambda_m\varphi_m^2 & \varphi_m^3 \end{bmatrix} \tag{19.31}$$

This least-squares solution is due to Wu and Yang (1981), with a fuller derivation by Snyder (1985), who has also developed corresponding matrices using complex algebra. Brief mention was made in Chapter 18 of the work of Rauhala (1972) and Kratky (1975) on speeding up the execution of least-squares solutions to determine the polynomial coefficient to transform Landsat MSS imagery. As in most other applications the numerical methods are equally suitable for dealing with any transformation of the type expressed by (19.12). In a conventional solution of this task the parameters are defined as components of a vector which is ultimately derived from a matrix equation system with the use of the least-squares adjustment. It was demonstrated by Rauhala (1972) that advantage can be taken of the symmetrical structure of certain two-dimensional polynomial expressions by the method of grouping the parameters in a two-dimensional array. The object is to keep the size of the matrices requiring inversion in (19.30) significantly smaller. Consequently the technique saves computer time. Rauhala's method was successfully used by the National Research Council of Canada for handling Landsat scenes and Kratky (1975) has demonstrated huge savings in computing time ranging from the ratio 1/3·5 for a second-degree polynomial through 1/23·5 for a seventh-degree polynomial, these being the comparison between calculations done by Rauhala's method and conventional solution.

As in the use of polynomial solutions for registration of scanned images, the efficacy and accuracy of numerical rather than analytical solutions

depends upon the size of the area mapped, and therefore the homogeneity of the data file. In making transformations of data originally digitised from paper maps the file may be heterogeneous simply because the positions of points may have been affected differently in different parts of the map by paper deformation and folding. Just as it is necessary to treat separately with the panels of a map which has, at some time, been folded, it may be necessary to divide the whole map into blocks and transform each block separately.

Solutions by interpolation methods

An important branch of numerical analysis is the creation of polynomial coefficients by interpolation. Once more these are techniques used for a multitude of other scientific applications.

There are two major kinds of solution; those for which the dependent variable is equally spaced and those for which it is not.

A typical example is that to be found in almost every elementary text on numerical methods. This is to find by interpolation a particular value of sin x using a table formed from equally spaced values of x.

The second type of interpolation arises more commonly in experimental work when observed values of a variable x do not occur at conveniently regular intervals.

Both of these methods have been employed in the study of map projections and for transformations from one surface to another. The first was much used by Ginzburg and Salmanova (1962, 1964) to create projections for small-scale world maps. The series of world polyconic projections created by interpolation were listed as Ginzburg IV through Ginzburg VII in Maling (1960).

Application of finite element interpolation

The reader who is aware of modern developments in civil engineering will know about the Finite Element Method, as described, and largely pioneered by Zienkiewicz (1977), which is used in design, structural analysis and many more applications. The finite element method is especially useful for making two dimensional transformations and would naturally serve very well as a tool in remote sensing and GIS. Indeed the application of Lagrangian interpolation in two dimensions has already been employed successfully by Spiess and Brandenberger (1989) for small-scale cartography. Obviously the method has enormous applications in other fields of conventional mapping as well as GIS. An important advantage is that there is now available a considerable amount of computer software.

TABLE 19.01 *Table showing the form of notation used in interpolation by divided differences*

x	$f(x)$	First differences	Second differences	Third differences
x_0	$f(x_0)$			
		$f(x_1, x_0)$		
x_1	$f(x_1)$		$f(x_2, x_1, x_0)$	
		$f(x_2, x_1)$		$f(x_3, x_2, x_1, x_0)$
x_2	$f(x_2)$		$f(x_3, x_2, x_1)$	
		$f(x_3, x_2)$		$f(x_4, x_3, x_2, x_1)$
x_3	$f(x_3)$		$f(x_4, x_3, x_2)$	
		$f(x_4, x_3)$		
x_4	$f(x_4)$			

Interpolation by divided differences

The method of interpolation to be employed in this context where the variables are not equally spaced was originally described by Newton, and it has been used by Lauf and Young (1961) to transform from one conformal projection to another.

If we have a series of values of x to which there correspond values $f(x)$, then we may construct the typical table. In Table 19.01 the first differences entries have the meaning

$$f(x_1, x_0) = [f(x_1) - f(x_0)]/[x_1 - x_0] \tag{19.32}$$

$$f(x_2, x_1) = [f(x_2) - f(x_1)]/[x_2 - x_1] \tag{19.33}$$

etc., the second differences entries correspond to

$$f(x_2, x_1, x_0) = [f(x_2, x_1) - f(x_1, x_0)]/[x_2, -x_1] \tag{19.34}$$

etc. and the third differences correspond to

$$f(x_3, x_2, x_1, x_0) = [f(x_3, x_2, x_1) - f(x_2, x_1, x_0)]/[x_3 - x_2] \tag{19.35}$$

The system may be extended to incorporate higher-order differences, but, as Lauf has shown, third-order differences are usually sufficient. Generalising this result to the case of $f(x)$ being a polynomial of order n, we may write for Newton's formula

$$f(x) = f(x_0) + (x - x_0)f(x_1, x_0) + \ldots + (x - x_0)(x - x_1)\ldots (x - x_{n-1})f(x_n, x_{n-1}, \ldots, x_0) \tag{19.36}$$

where $f(x)$ is the variable to be determined by interpolation to correspond to the known value of x.

Lauf applied this method to the transformation of conformal projections using complex numbers. From equation (19.15) he put

$$z = x + iy \tag{19.37}$$

$$F(z) = Z = X + iY \tag{19.38}$$

and completed tables for z (corresponding to x) and Z (corresponding to $f(x)$) in Table 19.01. In this notation the first differences read

$$[Z_1, Z_2] = [Z_1 - Z_2]/[z_1 - z_2] \tag{19.39}$$

and the second differences are

$$[Z_1, Z_2, Z_3] = \{[Z_1 Z_2] - [Z_2 Z_3]/[z_1 - z_3]\} \tag{19.40}$$

Finally the value for Z, to be obtained by interpolation may be determined by repeated substitution, leading to

$$Z = Z_1 + (z - z_1)[ZZ_1] \tag{19.41}$$

$$Z = Z_1 + (z - z_1)[Z_1 Z_2] + (z - z_1)(z - z_2)[Z_1 Z_2 Z_3] \tag{19.42}$$

$$Z = Z_1 + (z - z_1)[Z_1 Z_2] + (z - z_1)(z - z_2)[Z_1 Z_2 Z_3] + \ldots + (z - z_1)(z - z_2)\ldots(z - z_{n-1})[Z_1 Z_2 \ldots Z_n] + R_n(Z) \tag{19.43}$$

where

$$R_n(Z) = (z - z_1)(z - z_2)\ldots(z - z_n)[ZZ_1 Z_2 \ldots Z_n] \tag{19.44}$$

Lauf and Young provide four worked examples in their 1961 paper. Two of these transform from Mercator's projection to the equatorial aspect Stereographic projections using different distributions of control points common to both systems. The third example makes the transformation from the 2° Transverse Mercator belts employed for cadastral mapping in South Africa into the UTM. The fourth example transforms from Lambert Conformal Conical into another version of the Transverse Mercator projection.

Vincenty (1987) has reported on this method which he used to transform from NAD 27 into NAD 83. He confirms that the method gives the same results as a polynomial based upon equations (19.25) and (19.26). The method is mathematically neat and it requires little programming effort. It is also very fast because it does not form or solve any normal equations. However, by its very nature it does not produce residuals, but gives the transformed coordinates directly. This can be an operational disadvantage, for if there is a blunder in input data, as in transferring typewritten material to the database, the method will not detect it. For this reason it may give unreliable answers when distortions are large, but this is not a unique feature of the method. Indeed, a gross error may be detected more easily than in a conventional least-squares solution where errors are distributed and therefore less easy to detect.

Least-squares collocation

The final form of numerical solution to be mentioned here is a comparatively new mathematical tool developed independently by Krarup

(1969) and Moritz (1980b). It has evolved from statistical methods to interpolate gravity anomalies, but it has been shown by Mikhail (1976) and Ruffhead (1987) to have wider applications, particularly in the present case of two-dimensional transformation. We shall see that the method is much more complicated than those already described and, as Vincenty (1987) has found, the accuracy is no better than the use of a polynomial transformation.

The objects of collocation can be stated in terms of vectors defined as follows:

- **x** is the vector of measurements,
- **n** is the vector of *noise component in the measurements*,
- **s′** is the vector of signal components in the measurements,
- **s** is the vector of signals which we wish to predict,
- **u** is the vector of unknown parameters which define the mathematical model.

Although this terminology is unfamiliar in the mapping sciences we may interpret the noise components as being the random or accidental errors of measurement. As in most adjustment and transformation procedures the objective is to distinguish **s′** from **n** and therefore find **s**.

In the most general case

$$\mathbf{x} = \mathbf{Au} + \mathbf{s}' + \mathbf{n} \tag{19.45}$$

where **A** is the design matrix arising from the constituent parts of the mathematical model.

Converting these rather esoteric ideas into the practical task of transforming from one grid to another, we have the two coordinates systems (x, y) and (X, Y) as before. In the first coordinate system x_i and y_i are known for $i = 1 \ldots n$ data points and $i = n+1 \ldots n+p$ unknown or computation points. The differences between the coordinates are ΔX_i and ΔY_i, where

$$\Delta X_i = x_i - X_i \tag{19.46}$$

$$\Delta Y_i = y_i - Y_i \tag{19.47}$$

Of course, these are known for the data points and we want to *predict* ΔX_i and ΔY_i at each of the unknown computation points. Equation (19.45) can be formed by treating the known shifts as the vector of the measurements and writing the equations for the n data points

$$\Delta X_i = u_1 + kX_i u_3 + kY_i u_4 + s' + n \tag{19.48}$$

$$\Delta Y_i = u_2 + kY_i u_3 + kX_i u_4 + s' + n \tag{19.49}$$

The terms in u constitute the mathematical model, which is equivalent to a translation, scaling and rotation, and k is a constant. The remaining

terms may be called 'signal' and 'noise', although they represent the correlated and uncorrelated components of the 'unmodelled' part of the shift. Methods of solving the expressions are not given here, but may be found in the literature cited.

As an experimental test of the method, Ruffhead (1987) used it to transform a sample of Ordnance Survey trigonometric points in England and Wales using 14 data points and 15 'unknown' points, from the datum OSGB36 into OSGB70(SN). The objective was to predict the coordinate shifts at the 15 points as if they were unknown and then compare these 'predicted' OSGB70(SN) coordinates with the actual values calculated by the Ordnance Survey. The result was a mean distance error of 31 cm for the computation points. Vincenty (1987) used the same OS data with a fifth-order complex polynomial developed by him for making the transformations in the North American datum and obtained marginally better results, corresponding to the mean distance errors of 22–25 cm depending upon how the residuals were treated.

APPENDIX I

Algebraic expressions for the coordinates and particular scales of the most important map projections

This Appendix gives the general functional expressions needed to determine the coordinates and distortion characteristics of each of the named classes of projections, together with a list of specific equations for particular members of each class. The list is organised according to the classification system described in Chapter 7 and illustrated by Table 7.02, p. 148. However, this list differs from the classification system by including certain modified versions of some projections which are already well known, and by incorporating some important transverse and oblique aspect projections.

Because of the importance of the members of Tobler's Group D, the order of presentation is as follows:

Group D: Cylindrical projections, Azimuthal projections, Conical projections.
Group C: Pseudocylindrical projections, Pseudoazimuthal projections, Pseudoconical projections.
Group A: Polyconic projections.

Note that we do not describe any map projections from Group B. As noted in Chapter 7, these have very little practical value in cartography. All the coordinate expressions given below *have been derived for a sphere of unit radius*. In order to obtain master grid coordinates to construct a graticule to a required scale, it is sufficient to multiply the numerical values of the coordinates for each graticule intersection by the appropriate value of r from Table 8.02, p. 162.

The most important modern systematic account in the English language of the common map projections in use is that by Snyder (1987a), which provides in about 350 pages a comprehensive statement about the derivation, history and use of most of those listed below, with samples of the calculations to be made with reference to these projections based upon

the spherical and spheroidal assumptions. Snyder and Voxland (1989) also provide a simple description of each projection, together with an illustration of the world graticule. The projections in the following list which are marked with an asterisk (*) are illustrated in Snyder and Voxland (1989).

Group D: Cylindrical class

General expressions for normal aspect cylindrical projections:

$$x = \lambda; \quad y = f(\varphi); \quad h = dy/d\varphi; \quad k = \sec\varphi; \quad p = hk;$$
$$\varepsilon = 90° - \theta' = 0$$

or, where modified

$$x = \cos\varphi_0 . \lambda; \quad y = f(\varphi); \quad h = dy/d\varphi; \quad k = \cos\varphi_0/\cos\varphi; \quad p = hk;$$
$$\varepsilon = 90° - \theta' = 0$$

Decreasing separation of parallels or sine series

1* *Cylindrical equal-area projection (Lambert).* First described in 1772.

Equal-area $x = \lambda \quad y = \sin\varphi \quad k = \sec\varphi \quad h = \cos\varphi$

See Table 6.01, p. 111, Figs 6.04, p. 112, 7.02, p. 130 and 7.03, p. 131.

1a *Modified cylindrical equal-area projection.* First described in 1848, various versions with different standard parallels described thereafter; usually in ignorance of what had already been done.

The following list is of named projections which differ only in the choice of the standard parallels, φ_0

	Date	*Author and/or name*	*Standard parallels*
1a1*	1855	Gall's orthographic projection	45°N and S
1a2*	1910	Behrmann's projection	30°
1a3	1929	Limiting case of hyperbolic equal-area (Craster)	37°04′
1a4	1935	Balthasart	50°
1a5	1947	§Trystan Edwards' projection	37°23′
1a6	1973	§Peters' orthogonal map of the world (incorrect version of 1a7)	46°02′
1a7*	1975	Peters' (intended version of map and therefore identical to Gall's orthographic)	45°

In this list projections marked § are not equal-area.

Equidistant spacing of parallels

2* *Cylindrical equidistant or Plate Carrée projection.* Sometimes attributed to Anaximander, *c.* 550 BC, but more likely to have first been described some centuries later, probably by Eratosthenes (*c.* 275–195 BC). For an interesting correspondence on the subject, and how this belief has been perpetuated, see letters by Robinson, Sharples and Maling in the Newsletter of the Computer Centre, University College London; issues for 1984 and 1985.

Equidistant $x = \lambda$; $y = \varphi$; $h = 1{\cdot}0$; $k = \sec\varphi$

2a* *Cassini-Soldner or Cassini's projection.* First described by Cassini de Thury in 1745. The transverse aspect of the Plate Carrée projection which was much used as the base for topographical and cadastral mapping until the 1930s. See Chapter 15, pp. 310–335 for a detailed description.

2b* *Modified cylindrical equidistant projection.* The version of the Plate Carrée projection having a pair of standard parallels. The best known is that with $\varphi_0 = 45°$. This was attributed by Ptolemy to Marinus of Tyre, *c.* AD 100, and is therefore known as Marinus' projection. It was independently rediscovered by Gall in 1855 and called by him *Gall's isographic projection.*

$x = \cos\varphi_0 . \lambda$; $y = \varphi$; $h = 1{\cdot}0$; $k = \cos\varphi_0/\cos\varphi$

Increasing separation of parallels or tangent series

3* *Mercator's projection.* Now known to have been used in China as the base for the Tunhuang star chart as early as 940 by Ch'ien Lo-Chih. This is described and illustrated in Ronan (1983), *The Cambridge Illustrated History of the World's Science.* In Europe it is supposed to have been used first by Etzlaub in 1511 and Mercator in 1569. The navigation applications were first described in detail by Wright (1599).

Conformal $x = \lambda$; $y = \ln\tan(\pi/4 + \varphi/2)$; $h = k = \sec\varphi$

Has the important additional property that all rhumb-lines are rectilinear. See Chapter 10, pp. 209–217, Fig. 10.10.

3a *Modified Mercator's projection.* Versions of Mercator's projection having a standard parallel in latitude φ_0. Most nautical charts are based upon a parallel φ_0 near the centre of the chart.

$x = \cos\varphi_0 . \lambda$; $y = \cos\varphi_0 . \ln\tan(\pi/4 + \varphi/2)$; $h = k = \cos\varphi_0/\cos\varphi$

3b* *Transverse Mercator projection.* First described for the sphere by Lambert in 1772. See Chapter 16, pp. 336–363 for a detailed description.

3c* *Oblique Mercator projection.* Various versions of skew oblique aspect conformal cylindrical projections have been used for topographical and cadastral mapping. These are not described in this book apart from a brief mention of the *Hotine Oblique Mercator* (HOM) or *Rectified Skew Orthomorphic* projection for mapping the ground tracks of satellites.

4a and **4b*** *Miller's cylindrical projections.* First described by O. M. Miller in 1942.

$$x = \lambda; \quad y = n \ln \tan(\pi/4 + \varphi/2m)$$

where n and m are constants. Two versions were described:

$$y = (5/4) \ln \tan(\pi/4 + 2\varphi/5); \quad y = (3/2) \ln \tan(\pi/4 + \varphi/3)$$

5* *Perspective cylindrical projection (Braun).* First described in 1867.

$$x = \lambda; \quad y = 2 . \tan(\varphi/2)$$

The parent projection which is less well known than the following modifications of it:

5a* *Gall's Stereographic projection.* First described in 1855. A modification of Braun's projection (5) with standard parallels in latitudes 45°N and S.

$$x = \cos\varphi_0 . \lambda = (\sqrt{2}/2) . \lambda; \quad h = (\sqrt{2}+2)/4 \sec^2(\varphi/2);$$

$$y = (\sqrt{2}+2)/2 \tan(\varphi/2); \quad k = (\sqrt{2}/2) \sec\varphi$$

5b *BSAM projection.* First described in 1937. Used for maps in the *Bolshoi Sovietskii Atlas Mira* (Great Soviet World Atlas). The version of Braun's projection (5) with standard parallels in latitudes 30°N and S.

Group D: Azimuthal class

General expressions for normal aspect azimuthal projections:

$$r = f_1(\chi) = F_1(\varphi); \quad \theta = \lambda; \quad x = r . \sin\theta; \quad y = r . \cos\theta;$$

$$h = -(dr/d\varphi) = (dr/d\chi); \quad k = r/\cos\varphi = r/\sin\chi; \quad \varepsilon = 90° - \theta' = 0$$

Decreasing separation of parallels

6* *Stereographic projection.* Attributed to Hipparchus, 160–125 BC.

Conformal $r = 2\tan(\chi/2); \quad \theta = \lambda; \quad h = \sec^2\chi; \quad k = \sec\chi$

See Fig. 1.07, p. 15.

7* *Gnomonic projection.* Known before 600 BC, used by Thales (636?–546? BC) for star maps. All great circles are rectilinear.

Orthodromic $r = \tan\chi; \quad \theta = \lambda; \quad h = \sec^2\chi; \quad k = \sec\chi$

8* *Minimum-error azimuthal projection (Airy).* First described in 1861.

Minimum-error $r = 2\,.\cot(\chi/2)\,.\ln\sec(\chi/2) + \tan(\chi/2); \quad \theta = \lambda$

9 *Breusing's (Geometric Mean) azimuthal projection.* First described in 1892. A combined projection which is the geometric mean of the Stereographic (6) and Azimuthal equal-area (12) projections

$r = \{2[\tan(\chi/2)\,.\sin(\chi/2)]\}^{1/2}; \quad \theta = \lambda$

10 *Breusing's (Harmonic Mean) azimuthal projection.* First described in 1892. A combined projection which is the harmonic mean of the Stereographic (6) and Azimuthal equal-area (12) projections. Practically indistinguishable from (8).

$r = 4\tan(\chi/4); \quad \theta = \lambda; \quad h = \sec^2\chi/4; \quad k = \sec(\chi/2)\sec^2(\chi/4)$

Equidistant spacing of parallels

11* *Azimuthal equidistant projection (Postel).* It has been claimed that the oldest celestial map on this projection is that of Conrad of Dyffenbach (1426).

Equidistant $r = \chi; \quad \theta = \lambda; \quad h = 1{\cdot}0; \quad k = \chi/\sin\chi$

See Fig. 7.01, p. 127.

Increasing separation of parallels

12* *Azimuthal equal-area projection (Lambert).* First described in 1772. See Chapter 10, pp. 196–201 and Figs 10.02, 10.03 and 10.04.

Equal-area $r = 2\,.\sin(\chi/2); \quad \theta = \lambda; \quad h = \cos(\chi/2); \quad k = \sec(\chi/2)$

13* *Orthographic projection.* Attributed to Apollonius, *c.* 240 BC; used by Hipparchus (160–125 BC).

$r = \sin\chi; \quad \theta = \lambda; \quad h = \cos\chi; \quad k = 1{\cdot}0$

Group D: Conical class

General expressions for normal aspect conical projections:

$r = f_1(\chi) = F_0(\varphi); \quad \theta = n\,.\,\lambda; \quad x = r\,.\sin\theta;$

$r = C - r\cos\theta$; $C = \text{const.}$; $n = \text{const.}$;

$h = -(dr/d\varphi)$; $k = nr/\cos\varphi$; $\varepsilon = 90° - \theta' = 0$

Decreasing separation of parallels

14 *Conical equal-area projection with one standard parallel.* Truncated pole. First described in 1772.

Equal-area $r = \{[(1+n^2)/n] - [2 . \cos\chi/n]\}^{1/2}$; $n = \cos\chi_0 = \sin\varphi_0$;

$\theta = n . \lambda$; $h = \sin\chi/n . r$; $k = n . r/\sin\chi$

14a* *Conical equal-area projection with two standard parallels (Albers).* Truncated pole. First described in 1805.

Equal-area $r = [C^2 + (4/n)\sin^2(\chi/2)]^{1/2}$; $n = [\cos\chi_1 + \cos\chi_2]/2$;

$\theta = n . \lambda$; $C = (2/n)\sin^2(\chi_2/2) . \sin^2(\chi_2/2)$;

$h = \sin\chi/n . r$; $k = n . r/\sin\chi$

15* *Conical equal-area projection with one standard parallel (Lambert).* Point pole. First described in 1772.

Equal-area $r = (2/\sqrt{n}) . \sin(\chi/2)$; $\theta = n . \lambda$; $n = \cos^2(\chi_0/2)$;

$h = [\cos\chi/2]/[\cos\chi_0/2]$; $k = [\cos\chi_0/2]/[\cos\chi/2]$

Equidistant spacing of parallels

16* *Equidistant conical projection with one standard parallel (Ptolemy).* Attributed to Ptolemy, AD 130. Truncated pole. See Chapter 10, pp. 202–207 for derivation in terms of φ. Illustrated in Fig. 10.07.

Equidistant $r = \tan\chi_0 + (\chi - \chi_0)$; $n = \cos\chi_0$;

$\theta = n . \lambda$; $h = 1{\cdot}0$; $k = n . r/\sin\chi$

16a* *Conical equidistant projection with two standard parallels (de l'Isle).* First described in 1745. Truncated pole. See Chapter 10, pp. 207–209 for derivation in terms of φ, and Fig. 10.08, p. 210, for illustration.

Equidistant $r = (1/n)\sin[(\chi_1 + \chi_2)/2]\sin[(\chi_1 - \chi_2)/2] + \chi$;

$n = [\cos(\chi_1 + \chi_2)/2 . \sin(\chi_1 - \chi_2)/2][(\chi_1 - \chi_2)/2]$;

$\theta = n . \lambda$; $h = 1{\cdot}0$; $k = n . r/\sin\chi$

17 *Conical equidistant projection with one standard parallel (Mendeleev).* First described in 1907. Point pole.

Equidistant $r = \chi$; $n = \sin(\chi_0/\chi)$;

$\theta = n . \lambda$; $h = 1{\cdot}0$; $k = n . r/\sin\chi$

18 *Conical projection (Murdoch I).* First described in 1758. Very close to the minimum-error conical projection (19)

$$r = m + \chi; \quad n = \cos(\chi_N \chi + \chi_S)/2; \quad \theta = n \,.\, \lambda;$$

$$m = \tan[(\chi_N + \chi_S)2] \,.\, \{\sin[(\chi_S - \chi_N)/2]/[(\chi_S - \chi_N)/2] - [(\chi_N + \chi_S)/2]\}$$

19 *Minimum-error conical projection (Murdoch III).* First described in 1758 but that version contained errors. The true minimum-error projection to satisfy Murdoch's theoretical specification was not described until 1904 by Everett. See Young (1923) and Maling (1983).

Minimum-error

$$\mathrm{r} = m + \chi; \; \theta = n \,.\, \lambda;$$

$$n = [\sin \tfrac{1}{2}(\chi_S - \chi_N)]/[\tfrac{1}{2}(\chi_S - \chi_N)] \,.\, [\sin \tfrac{1}{2}(\chi_S + \chi_N)]/[m + \tfrac{1}{2}(\chi_S + \chi_N)];$$

$$m = \tan \tfrac{1}{2}(\chi_S + \chi_N)[\tfrac{1}{2}(\chi_S - \chi_N) \cot \tfrac{1}{2}(\chi_S - \chi_N)]$$

Increasing separation of parallels

20a* *Conformal conical projection with one standard parallel (Lambert).* First described in 1772.

Conformal

$$\mathrm{r} = \tan \chi_0 [\tan \tfrac{1}{2}\chi / \tan \tfrac{1}{2}\chi_0]^n; \quad n = \cos \chi_0;$$

$$\theta = n \,.\, \lambda; \quad h = k = [\sin \chi_0 / \tan \tfrac{1}{2}\chi_0] \,.\, [\tan^n \tfrac{1}{2}\chi / \sin \chi]$$

20b* *Conformal conical projection with two standard parallels (Lambert).* First described in 1772.

Conformal

$$\mathrm{r} = \sin \chi_1 / n [\tan \tfrac{1}{2}\chi / \tan \tfrac{1}{2}\chi_1]^n;$$

$$n = [\ln \sin \chi_1 - \ln \sin \chi_2]/[\ln \tan \tfrac{1}{2}\chi_1 - \ln \tan \tfrac{1}{2}\chi_2];$$

$$\theta = n \,.\, \lambda; \quad h = k = n \,.\, \mathrm{r}/\sin \chi$$

Group C: Pseudocylindrical projections

General expressions for normal aspect pseudocylindrical projections:

$$\mathrm{x} = f_1(\varphi, \lambda); \quad \mathrm{y} = f_2(\varphi); \quad h = \partial \mathrm{y}/\partial \varphi \sec \varepsilon;$$

$$k = \partial \mathrm{x}/\partial \lambda \sec \varphi; \quad p = h \,.\, k \,.\, \cos \varepsilon$$

where $\varepsilon = 90° - \theta°$, ψ is an auxiliary angle which is a function of latitude usually expressed by a transcendental equation. This has to be solved by the Newton–Ralphson or 'Regula falsi' methods of numerical analysis, although graphical solutions are sometimes used. Because $\varepsilon > 0$ there are *neither conformal nor equidistant members of this class.*

Decreasing separation of parallels or sine series

21* *Mollweide's projection.* First described in 1805. Elliptical meridians. See Fig. 6.07, p. 117.

Equal-area $x = [\sqrt{8}/\pi] . \lambda . \cos\psi$; $y = \sqrt{2} . \sin\psi$

where ψ is the auxiliary angle to be found from the transcendental equation

$$\sin 2\psi - 2\psi = \pi . \sin\varphi; \quad h = \sec\varepsilon/k; \quad k = (2\sqrt{2}\cos\psi)/\pi . \cos\varphi$$

22 *Pseudocylindrical equal-area projection with elliptical meridians (Fournier II).* First described in 1646.

Equal-area $x = n . \lambda . \cos\varphi$; $y = n . \pi/2 . \sin\varphi$; $n = 1/\sqrt{\pi}$

23* *Pseudocylindrical equal-area projection with elliptical meridians and pole-line (Eckert IV).* First described in 1906.

Equal-area $x = [0{\cdot}84447\lambda/2](1+\cos\psi)$; $y = [0{\cdot}84447\pi/2]\sin\psi$

where ψ is the auxiliary angle to be found from the transcendental equation

$$2\psi + 4\sin\psi + \sin 2\psi = (4+\pi)\sin\varphi$$

24* *Parabolic projection (Craster).* First described in 1929. Parabolic meridians.

Equal-area $x = \lambda(3/\pi)^{1/2}[1-(4y^2/3\pi)]$; $y = \sqrt{3\pi} . \sin(\varphi/3)$

25* *Pseudocylindrical equal-area projection with sinusoidal meridians and pole-line (Eckert VI).* First described in 1906. See Fig. 13.05, p. 276 for a recentred version.

Equal-area $x = (0{\cdot}882\lambda/2)\cos^2(\psi/2)$; $y = 0{\cdot}882\psi$

ψ is the auxiliary angle to be obtained from the transcendental equation

$$\psi + \sin\psi = (\lambda/0{\cdot}882^2)\sin\varphi$$

26 *Pseudocylindrical equal-area projection with sinusoidal meridians and pole-line (Nell–Hammer).* First described in 1890. A combined projection which is the arithmetic mean of the coordinates for the Cylindrical equal-area projection (1) and the Sinusoidal projection (30)

Equal-area $x = \frac{1}{2}\lambda(1+\cos\varphi)$; $y = \frac{1}{2}(\varphi+\sin\varphi)$

27 *Pseudocylindrical equal-area projection with sinusoidal meridians and pole-line (Kavraisky V).* First described in 1933.

Equal-area $x = (1/mn)\lambda \,.\, \sec n\varphi \,.\, \cos\varphi$; $y = m \,.\, \sin n\varphi$;

$m = 1{\cdot}504875$; $n = 0{\cdot}738341$

28 *Pseudocylindrical equal-area projection with sinusoidal meridians and pole-line (Kavraisky VI).* First described in 1936.

Equal-area $x = 0{\cdot}877\lambda \,.\, \cos\psi$; $y = 1{\cdot}3161\psi$; $\sin\psi = (\sqrt{3}/2)\sin\varphi$

29* *Pseudocylindrical equal-area projections (Boggs) or 'Boggs Eumorphic projection'.* First described in 1929; a combined projection whose coordinates are the arithmetic mean of Mollweide's projection (21) and the Sinusoidal projection (30).

Equal-area $x = \frac{1}{2}\lambda[(2\sqrt{2}/\pi)\cos\psi + \cos\varphi]$; $y = \frac{1}{2}(\varphi + \sqrt{2}\sin\psi)$

where ψ was defined for Mollweide's projection (21).

Equidistant spacing of parallels

30* *Sinusoidal or Sansom–Flamsteed projection.* First described in 1606. See Fig. 6.06, p. 115, Fig. 7.04, pp. 132–133 and Fig. 15.01, p. 314.

Equal-area $x = \lambda \,.\, \cos\varphi$; $y = \varphi$; $h = \sec\varepsilon$; $k = \cos\varepsilon$

30a *Modified Sinusoidal projection (Tissot).* First described in 1881.

Equal-area $x = n \,.\, \lambda \,.\, \cos\varphi$; $y = m \,.\, \varphi$; $h = m \,.\, \sec\varepsilon$;

$k = n \,.\, \cos\varepsilon$; $m = 0{\cdot}875$; $n = 1{\cdot}25$

31* *Pseudocylindrical projection with elliptical meridians (Apianus II).* First described in 1524.

$x = \lambda \,.\, \cos\psi$; $y = (\pi/2) \,.\, \sin\psi$

where:

$\sin\psi = 2\varphi/\pi$; $h = \sec\varepsilon$; $k = \cos\psi/\cos\varphi$

32* *Pseudocylindrical projections with elliptical meridians and pole-line (Eckert III or Ortelius' projection).* First described in 1570. A combined projection which is the arithmetic mean of the Plate Carrée (2) and Apianus II (31)

$x = \frac{1}{2}(0{\cdot}84\lambda) \,.\, [1 + \cos\psi]$; $y = \frac{1}{2}(0{\cdot}844\pi)\sin\psi$; $\sin\psi = 2\varphi/\pi$

33 *Pseudocylindrical projection with sinusoidal meridians and pole-line (Eckert V).* First described in 1906.

$$x = \tfrac{1}{2}m \,.\, \lambda(1+\cos\varphi); \quad y = m \,.\, \varphi; \quad m = 2/[(\pi+2)^{1/2}] = 0{\cdot}822\ldots$$

Increasing separation of parallels or tangent series

A few pseudocylindrical projections classified within this series have been described by Maurer and van der Grinten, but none of them has any practical value.

Group C: Pseudoconical class

General expressions for normal aspect pseudoconical projections.

$$r = f_1(\chi) = F_1(\varphi); \quad \theta = f_2(\varphi, \lambda); \quad x = \sin\theta; \quad y = q - r \,.\, \cos\theta;$$

$$\tan\varepsilon = [r(\partial\theta/\partial\varphi)]/(dr/d\varphi); \quad h = -(dr/d\varphi)\sec\varphi;$$

$$k = [r/\cos\varphi] \,.\, [\theta/\partial\lambda]; \quad q = \text{const}; \quad p = h \,.\, k\cos\varepsilon$$

Because $\varepsilon > 0$ *there are neither conformal nor equidistant members of this class.*

Equidistant spacing of parallels

35* *Bonne's* projection. First described in 1520.

Equal-area $r = (\cot\varphi_0 + \varphi_0) - \varphi; \quad \tan\varepsilon = \lambda \,.\, \sin\varphi - \theta = 2\tan(\omega/2);$

$$\theta = [\cos\varphi/r] \,.\, \lambda; \quad h = \sec\varepsilon; \quad k = p = 1{\cdot}0$$

35a* *Stab-Werner projection.* First described in 1514. The limiting case of (35) where $\varphi_0 = 90°$.

Group C: Pseudoazimuthal class

General expressions for normal aspect pseudoazimuthal projections:

$$r = f_1(\chi); \quad \theta = f_2(\chi, \lambda); \quad x = r \,.\, \sin\theta; \quad y = r \,.\, \sin\theta;$$

$$\tan\varepsilon = [r(\partial\theta/\partial\chi)]/[dr/d\chi]; \quad h = [dr/d\chi] \,.\, \sec\varepsilon;$$

$$k = [r/\sin\chi] \,.\, [\partial\theta/\partial\lambda]; \quad p = h \,.\, k \,.\, \cos\varepsilon$$

Because $\varepsilon > 0$ *there are neither conformal nor equidistant members of this class.*

Increasing separation of parallels

36* *Pseudoazimuthal equal-area projection (Wiechel).* First described in 1879.

Equal-area $r = 2\sin(\chi/2); \quad \theta = \lambda + (\chi/2)$

36a *TsNIIGAiK projection with oval isograms (Ginzburg III).* First described in 1952 in oblique aspect only.

$$r = 3\sin(z/3)$$

where z is the angular distance from the origin and z_β is the maximum extent of the map, and (z, α) are the bearing and distance coordinates of a point

$$\theta = \alpha - C[z/z_\beta] \,.\, \sin n \,.\, \alpha; \quad z_\beta = 120°; \quad C = 0{\cdot}15; \quad m = 2; \quad n = 2$$

Group A: Polyconic projections

General expressions for normal aspect polyconic projections:

$$x = f_1(\varphi, \lambda); \quad y = f_2(\varphi, \lambda); \quad r = f_3(\varphi, \lambda); \quad \theta = f_4(\varphi, \lambda)$$

Particular scales and distortion characteristics from basic formulae as shown in Chapter 6, pp. 118–121.

Decreasing separation of parallels

37* *Hammer–Aitoff projection.* First described in 1892.

Equal-area
$$x = 2\sqrt{2}[(\cos\varphi \,.\, \sin \tfrac{1}{2}\lambda)/(1 + \cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)^{1/2}];$$
$$y = \sqrt{2}[\sin\varphi/(1 + \cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)^{1/2}]$$

See Fig. 6.08, p. 119 and the derivation of equations (6.34)–(6.39) for the particular scales.

37a *Modified Hammer–Aitoff projection.* First described in 1953.

Equal-area
$$x = 2\sqrt{2} \,.\, m[(\cos\varphi \,.\, \sin \tfrac{1}{2}\lambda)/(1 + \cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)^{1/2}];$$
$$y = (\sqrt{2}/m) \,.\, [\sin\varphi/(1 + \cos\varphi \,.\, \cos \tfrac{1}{2}\lambda)^{1/2}]; \quad m = (a/2b)^{1/2}$$

where a and b are the semi-axes of the ellipse formed by the boundary of the world map.

37b* *Briesemeister's projection.* First described in 1952.

Equal-area Oblique aspect of (37a) where $\varphi_0 = 45°N$, $\lambda_0 = 15°E$ and $m = 0{\cdot}9354$.

See Fig. 8.02, p. 158, pp. 156–162 and Appendix II, p. 442.

37c *Bomford's projection.* First described in 1952. Skew oblique aspect of (37) with origin in $\varphi_0 = 45°N$; $\lambda_0 = 10°W$.

37d* *Hammer–Wagner projection.* First described in 1949. Version of Hammer–Aitoff projection (37) with curved pole-line.

Equal-area $x = 5{\cdot}33448 \sin(z/2) \sin\alpha;$

$y = 2{\cdot}48206 \sin(z/e) \cos\alpha; \quad \sin\psi = 0{\cdot}90632 \sin\varphi;$

$\cos z = \cos(\lambda/3) \cos\psi; \quad \cos\alpha = \sin\psi/\sin z$

Equidistant spacing of parallels

38* *The Polyconic projection or Simple Polyconic projection.* Apparently first described by Hassler about 1820. See Fig. 5.02, p. 86.

$x = \cot\varphi \,.\, \sin(\lambda \,.\, \sin\varphi); \quad y = \varphi + 2\cot\varphi \,.\, \sin^2[(\lambda \,.\, \sin\varphi)/2]$

38a* *Modified Polyconic projection for the International Map of the World at 1/1 000 000.* Not described in this book. See Snyder (1987a), pp. 131–137.

39* *Aitoff's projection.* First described in 1889.

$x = 2z \,.\, \sin\alpha; \quad y = z \,.\, \cos\alpha; \quad \cos z = \cos\varphi \,.\, \cos\tfrac{1}{2}\lambda$

$\cot\alpha = \tan\varphi \,.\, \operatorname{cosec}\tfrac{1}{2}\lambda; \quad \cos\alpha = \sin\lambda/\sin z$

39a* *Aitoff–Wagner projection.* First described in 1949. Modified from Aitoff's projection (39) with curved pole-line. Illustrated in Fig. 1.05, p. 8.

$x = 3{\cdot}6z \,.\, \sin\alpha; \quad y = 1{\cdot}28571z \,.\, \cos\alpha;$

$\cos z = \cos[5\lambda/16] \cos[7\varphi/9]; \quad \cos\alpha = [\sin(7\varphi/9)]/\sin z$

40* *Tripel projection (Winkel).* First described in 1913. A combined projection which is the arithmetic mean of the Plate Carrée (2) and Aitoff's (39) projection.

$x = \tfrac{1}{2}[n \,.\, \lambda + 2z \,.\, \sin\alpha]; \quad y = \tfrac{1}{2}[\varphi + z \,.\, \cos\alpha];$

$\cos z = \cos\tfrac{1}{2}\lambda \,.\, \cos\varphi; \quad \cos\alpha = \sin\varphi/\sin z; \quad n = \cos\varphi_0 = \cos 40^\circ$

APPENDIX II

Table of projection coordinates for 15° graticule intersections of Briesemeister's projection

Origin of projection: $\varphi_0 = 45°$ North, $\lambda_0 = 15°$ East. Coordinates for eastern hemisphere only; to obtain those for the western hemisphere, use negative values for x throughout.

Latitude (φ)	Longitude (λ) 15° E	30° E	45° E	60° E	75° E	90° E	105° E	120° E	135° E	150° E	165° E	180° E	165° W	
90° N	0.00000													x
	+0.81823													y
75° N	0.00000	+0.06499	+0.12627	+0.18018	+0.22327	+0.25261	+0.26552	+0.26039	+0.23668	+0.19535	+0.13913	+0.07228	0.00000	x
	+0.55338	+0.56093	+0.58319	+0.61910	+0.66687	+0.72407	+0.78773	+0.85421	+0.91938	+0.97851	+1.02644	+1.05800	+1.06906	y
60° N	0.00000	+0.12254	+0.23933	+0.34453	+0.43221	+0.49647	+0.53161	+0.53247	+0.49487	+0.41665	+0.29970	+0.15438	0.00000	x
	+0.27908	+0.29247	+0.33204	+0.39596	+0.48133	+0.58447	+0.70092	+0.82565	+0.95286	+1.07575	+1.18547	+1.26838	+1.30160	y
45° N	0.00000	+0.17211	+0.33779	+0.49021	+0.62195	+0.72519	+0.79193	+0.81418	+0.78416	+0.69428	+0.53741	+0.30733	0.00000	x
	0.00000	+0.01821	+0.07194	+0.15836	+0.27334	+0.41186	+0.56833	+0.73685	+0.91106	+1.08409	+1.24820	+1.39441	+1.51188	y
30° N	0.00000	+0.21298	+0.42013	+0.61451	+0.78786	+0.93122	+1.03555	+1.09232	+1.09385	+1.03449	+0.91474	+0.75943	+0.68478	x
	−0.27908	−0.25665	−0.19095	−0.08618	+0.05181	+0.21648	+0.40105	+0.59855	+0.80171	+1.00234	+1.19009	+1.34975	+1.46036	y
15° N	0.00000	+0.24439	+0.48500	+0.71535	+0.92635	+1.10819	+1.25173	+1.34950	+1.39649	+1.39208	+1.34916	+1.30163	+1.29674	x
	−0.55338	−0.52707	−0.45072	−0.33082	−0.17551	+0.00698	+0.20873	+0.42194	+0.63837	+0.84868	+1.04062	+1.19930	+1.35192	y
Equator	0.00000	+0.26552	+0.53161	+0.79192	+1.03556	+1.25173	+1.43189	+1.57037	+1.66533	+1.72093	+1.75152	+1.78697	+1.87085	x
	−0.81822	−0.78773	−0.70092	−0.56833	−0.40105	−0.20873	0.00000	+0.21691	+0.43332	+0.63931	+0.82303	+0.97074	+1.06906	y
15° S	0.00000	+0.27602	+0.56142	+0.84665	+1.11674	+1.36035	+1.57037	+1.74342	+1.88017	+1.98652	+2.07564	+2.16875	+2.29132	x
	−1.06906	−1.03255	−0.93341	−0.78995	−0.61629	−0.42250	−0.21691	−0.00757	+0.19710	+0.38735	+0.55213	+0.67913	+0.75593	y
30° S	0.00000	+0.27938	+0.58572	+0.89128	+1.17793	+1.43727	+1.66533	+1.86116	+2.02684	+2.16809	+2.29465	+2.41953	+2.55563	x
	−1.30160	−1.25130	−1.13417	−0.98287	−0.81053	−0.62505	−0.43332	−0.24245	−0.06027	+0.01441	+0.24175	+0.34102	+0.39130	y
45° S	0.00000	+0.33907	+0.66122	+0.96303	+1.24155	+1.49457	+1.72093	+1.92087	+2.09632	+2.25113	+2.39071	+2.52092	+2.64579	x
	−1.51188	−1.40187	−1.27213	−1.12647	−0.96916	−0.80498	−0.63931	−0.47818	−0.32835	−0.19732	−0.09323	−0.02463	0.00000	y
60° S	+0.68478	+0.75012	+0.92563	+1.13690	+1.35224	+1.55900	+1.75152	+1.92729	+2.08566	+2.22711	+2.35258	+2.46250	+2.55563	x
	−1.46036	−1.39656	−1.30486	−1.19476	−1.07395	−0.94827	−0.82303	−0.70355	−0.59538	−0.50431	−0.43622	−0.39685	−0.39130	y
75° S	+1.29674	+1.32892	+1.37483	+1.45703	+1.55976	+1.67235	+1.78697	+1.89811	+2.00190	+2.09551	+2.17655	+2.24271	+2.29132	x
	−1.24192	−1.27883	−1.23271	−1.17504	−1.10947	−1.04006	−0.97074	−0.90537	−0.84759	−0.80092	−0.76846	−0.75285	−0.75593	y
90° S	+1.87086													x
	−1.06906													y

APPENDIX III

The Transverse Mercator equations for a projection of the spheroid in detail

Eight equations are required to determine position, convergence and local scale-factor from both geographical coordinates and grid coordinates. These are listed in detail here rather than in Chapter 16, where they would interrupt the narrative unduly. Here they are given for two versions: the direct or Gauss–Krüger projection and the Biernacki–Rapp method of double projection. All the Gauss–Krüger solutions are provided, but the only expressions for double projection listed here are the expansions needed to determine the spherical coordinates (Φ, Λ) in double-projection. The remainder of that solution is given in the text of Chapter 16.

The Gauss–Krüger projection

The equations are listed here in two forms. The first entry for each solution is the direct method, comprising the terms in t and η to powers of λ^7 and λ^8 in longitude; E^6 and E^7 in Eastings. These are some of the famous equations published by Redfearn in 1948. As indicated in Chapter 16, pp. 361–363, use of these in a computer program is likely to create overflow conditions. Consequently one of the nested solutions is also given for each equation. This version was written by Meade (1987) for the UTM, but has been modified here to allow solutions for any scale factor k_0 on the central meridian. The notation in these equations is the same as that used in Redfearn's, with the additional parameter L to serve as the longitude term. We write

$$L = \lambda . \cos \varphi \qquad \text{(III.1)}$$

Note also that longitude is measured from the central meridian of the projection and therefore corresponds to the longitude terms $\delta\lambda$ or ω used by many authors (though not used in this book). The Eastings term, E, is similarly measured from the ordinate corresponding to the central meridian and no allowance has been made for any convention relating

the False Origin to the True Origin or for any other latitude of origin than the equator. In equations (III.28)–(III.47) the primed letters φ', ν', t', η' refer to the foot-point latitude, which must be determined in advance.

Input = geographical coordinates

Easting and Northing from (φ, λ)

Redfearn's Eastings equation:

$$\begin{aligned} E = k_0 . \nu[&\lambda . \cos\varphi + (\lambda^3/6)\cos^3\varphi(1 - t^2 + \eta^2) \\ &+ (\lambda^5/120)\cos^5\varphi(5 - 18t^2 + t^4 + 14\eta^2 + 13\eta^4 + 4\eta^6 - 58\eta^2 t^2 \\ &- 64\eta^4 t^2 - 24\eta^6 t)^2 + (\lambda^7/5040)\cos^7\varphi(61 - 479t^2 + 179t^4 - t^6)] \end{aligned} \quad \text{(III.2)}$$

Meade's nested solution for the Eastings equation:

$$E = A_1 L[1 + L^2(A_3 + L^2\{A_5 + A_7 L^2\})] \quad \text{(III.3)}$$

where

$$A_1 = k_0 . \nu \quad \text{(III.4)}$$

$$A_3 = (1 - t^2 + \eta^2)/6 \quad \text{(III.5)}$$

$$A_5 = [5 - t^2(18 - t^2) + \eta^2(14 - 58t^2)]/120 \quad \text{(III.6)}$$

$$A_7 = [61 - t^2(479 - 179t^2 + t^4)]/5040 \quad \text{(III.7)}$$

Redfearn's Northings equation:

$$\begin{aligned} N = k_0 m + k_0 . \nu[&(\lambda^2/2)\sin\varphi . \cos\varphi + (\lambda^4/24)\sin\varphi \\ &. \cos^3\varphi(5 - t^2 + 9\eta^2 + 4\eta^4) + (\lambda^6/720)\sin\varphi . \cos^5\varphi(61 - 58t^2 + t^4 \\ &+ 270\eta^2 + 445\eta^4 + 324\eta^4 + 88\eta^8 - 330\eta^2 t^2 - 680\eta^4 t^2 - 600\eta^6 t^2 \\ &- 192\eta^8 t^2) + (\lambda^8/40320)\sin\varphi . \cos^7\varphi(1385 - 3111t^2 + 543t^4 - t^6)] \end{aligned} \quad \text{(III.8)}$$

Meade's nested solution Northings:

$$N = k_0 . m + A_2 L^2[1 + L^2(A_4 + A_6 L^2)] \quad \text{(III.9)}$$

where

$$A_2 = \tfrac{1}{2} k_0 . \nu . t \quad \text{(III.10)}$$

$$A_4 = [5 - t^2 + \eta^2(9 + 4\eta^2)]/12 \quad \text{(III.11)}$$

$$A_6 = [61 - t^2(58 - t^2) + \eta^2(270 - 330t^2)]/360 \quad \text{(III.12)}$$

Convergence from φ and λ

Redfearn's convergence equation:

$$\gamma = k_0[\lambda . \sin\varphi + \lambda^3(1/3)\sin\varphi . \cos^2\varphi(1+3^2+2\eta^4) + \lambda^5(1/15)\sin\varphi . \cos^4\varphi(5-4t^2+14\eta^2+13\eta^4+4\eta^6-28\eta^2t^2 - 48\eta^4t^2-24\eta^6t^2) + \lambda^7(1/315)\sin\varphi . \cos^6\varphi(17-26t^2+2t^4)] \quad \text{(III.13)}$$

Meade's nested solution for convergence:

$$\gamma = C_1L[1+L^2(C_3+C_5L^2)] \quad \text{(III.14)}$$

$$C_1 = t \quad \text{(III.15)}$$

$$C_3 = [1+\eta^2(3+2\eta^2)]/3 \quad \text{(III.16)}$$

$$C_5 = (2-t^2)/15 \quad \text{(III.17)}$$

Local Scale-Factor from φ and λ

Redfearn's equation:

$$k = k_0[1+\lambda^2(1/2)\cos^2\varphi(1+\eta^2)+\lambda^4(1/24)\cos^4\varphi(5-4t^2 + 14\eta^2+13\eta^4+4\eta^6-28\eta^2t^2-48\eta^4t^2-24\eta^6t^2) + \lambda^6(1/720)\cos^6\varphi(61-148t^2+16t^4)] \quad \text{(III.18)}$$

Meade's nested solution for local scale-factor:

$$k = k_0[1+F_2L^2(1+F_4L^2)] \quad \text{(III.19)}$$

$$F_2 = (1+\eta^2)/2 \quad \text{(III.20)}$$

$$F_4 = [5-4t^2+\eta^2(9-24t^2)]/12 \quad \text{(III.21)}$$

Input = grid coordinates

The foot-point latitude

In order to make the grid-geographical transformation, we have to find a preliminary value for latitude from which ν, ρ, and the various trigonometric functions $\sin\varphi$, $\cos\varphi$, $\tan\varphi$ must be determined. The true value for the latitude is unknown (for this is what we are trying to determine); therefore an estimate has to be made for that latitude where the perpendicular through the point (E, N) meets the central meridian.

We have already called this the foot-point latitude in Chapter 15, and again we designate it as ϕ'. It is, of course, the latitude of the point M in Fig. 16.02 which has grid coordinates (0, N′) which is where the perpendicular from the point A meets the central meridian. In the days when the Transverse Mercator projection formulae were solved with the aid of tables, φ' was found by inverse interpolation in the tabulated values for m to find a value of latitude which corresponded to the value for m which was closest to the northing, N. Nowadays the foot-point latitude has to be computed. This is commonly done by an iterative method, such as has been described by Edoga (1981). However a one-step solution, by Sodano (1965), which has also been described by Makris (1982) is given here.

As a first approximation, we put

$$\varphi_0 = \mathrm{N}/\mathrm{b} \tag{III.22}$$

we may determine the reduced latitude, u, from

$$\mathrm{u} = \mathrm{E}\varphi_0 + \mathrm{F}\sin 2\varphi_0 + \mathrm{G}(5\sin 2\varphi_0 - 8\varphi_0)\cos^2\varphi_0 \tag{III.23}$$

where

$$\mathrm{E} = [1 - (1/4)\mathrm{e}'^2 + (11/64)\mathrm{e}'^4 \ldots] \tag{III.24}$$

$$\mathrm{F} = [(1/8)\mathrm{e}'^2 - (13/128)\mathrm{e}'^4] \tag{III.25}$$

$$\mathrm{G} = \mathrm{e}'^2/64 \tag{III.26}$$

Finally

$$\varphi' = \tan^{-1}[(\mathrm{a}/\mathrm{b})\tan \mathrm{u}] \tag{III.27}$$

where u is the reduced latitude corresponding to the required foot-point latitude; and a, b are the semi-axes of the spheroid.

Latitude and longitude from (E, N)

Redfearn's equation for latitude:

$$\begin{aligned}\varphi = \varphi' &- \mathrm{E}^2(\mathrm{t}'/2\mathrm{k}_0\rho'\nu')(1+\eta^2) \\ &+ \mathrm{E}^4(\mathrm{t}'/24\mathrm{k}_0\rho'\nu'^3)(5+3\mathrm{t}'^2+\eta^2-4\eta^4-9\eta^2\mathrm{t}'^2) \\ &- \mathrm{E}^6(\mathrm{t}'/720\mathrm{k}_0\rho'\nu'^5)(61+90\mathrm{t}'^2+45\mathrm{t}'^4+46\eta^2-3\eta^4+100\eta^6+88\eta^8 \\ &-252\eta^2\mathrm{t}'^2-66\eta^4\mathrm{t}'^2+84\eta^6\mathrm{t}'^2-192\eta^8\mathrm{t}'^2-90\eta^2\mathrm{t}'^4+225\eta^4\mathrm{t}'^4) \\ &+ \mathrm{E}^8(\mathrm{t}'/40320\mathrm{k}_0\rho\,.\,\nu'^7)(1385+3633\mathrm{t}'^2+4095\mathrm{t}'^4+1575\mathrm{t}'^6)\end{aligned} \tag{III.28}$$

Meade's nested solution for latitude:

$$\varphi = \varphi' + \mathrm{B}_2\mathrm{Q}^2[1+\mathrm{Q}^2(\mathrm{B}_4+\mathrm{B}_6\mathrm{Q}^2)] \tag{III.29}$$

$$\mathrm{Q} = \mathrm{E}/\mathrm{k}_0\,.\,\nu' \tag{III.30}$$

$$B_2 = -t'(1+\eta'^2)/2 \tag{III.31}$$

$$B_4 = -[5+3t'^2+\eta'^2(1-9t'^2-4\eta'^2)]/12 \tag{III.32}$$

$$B_6 = [61+t'^2(90+45t'^2)+\eta'^2(46-252t'^2-90t'^4)]/360 \tag{III.33}$$

Redfearn's equation for longitude:

$$\begin{aligned}\lambda = {} & E(\sec\varphi'/k_0v') - E^3(\sec\varphi'/6k_0v'^3)(1+t'^2-\eta^2-2\eta^4) \\ & + E^5(\sec\varphi'/120k_0v'^5)(5+28t'^2+24t'^4+6\eta^2-3\eta^4-4\eta^6 \\ & + 8\eta^2t'^2+4\eta^4t'+24\eta^6t'^2) - E^7(\sec\varphi'/5040k_0v^7)(61 \\ & + 662t'^2+1320t'^4+720t'^6)\end{aligned} \tag{III.34}$$

Meade's nested solution for longitude:

$$L = Q[1+Q^2(B_3+Q^2\{B_5+B_7Q^2\})] \tag{III.35}$$

$$B_3 = -(1+2t'^2+\eta'^2)/6 \tag{III.36}$$

$$B_5 = [5+t'^2(28+24t'^2)+\eta'^2(6+8t'^2)]/120 \tag{III.37}$$

$$B_7 = -[61+t'^2(662+1320t'^2+720t'^4)]/5040 \tag{III.38}$$

Convergence from (E, N)

Redfearn's equation for convergence:

$$\begin{aligned}\gamma = {} & E(t'/v')E^3(t'/3v'^3)(1+t'^2-\eta^2-2^4)+E^5(t'/15v'^5)(2+5t'^2+3t'^4 \\ & + 2\eta^2+9\eta^4+20\eta^6+11\eta^8+\eta^2t'^2-7\eta^4t'^2-26\eta^6t'^2-24\eta^8t'^2) \\ & - E^7(t'/315k_0v^7)(17+77t'^2+105t'^4+45t'^6)\end{aligned} \tag{III.39}$$

Meade's nested solution for convergence:

$$\gamma = D_1Q[1+Q^2(D_3+D_5Q^2)] \tag{III.40}$$

$$D_1 = t' \tag{III.41}$$

$$D_3 = -[1+t'^2-\eta'^2(1+2\eta'^2)]/3 \tag{III.42}$$

$$D_5 = [2+t'^2(5+3t'^2)]/15 \tag{III.43}$$

Local scale-factor from E, N

Redfearn's equation for local scale-factor:

$$\begin{aligned}k = k_0[1+E^2(1/2v'^2)(1+\eta^2)+E^4(1/24v'^4)(1+6\eta^2 \\ +9\eta^4+4\eta^6-24\eta^4t'^2-24\eta^6t'^2)+E^6(1/720v'^6)]\end{aligned} \tag{III.44}$$

Meade's nested solution for local scale-factor:

$$k = k_0[1+G_2Q^2(1+G_4Q^2)] \tag{III.45}$$

$$G_2 = (1+\eta'^2)/2 \tag{III.46}$$

$$G_4 = (1+5\eta'^2)/12 \tag{III.47}$$

Correction to a line of finite length

The scale-factor applies to an infinitely short line at one point so that if a line extends some distance from it in an East–West direction more than one scale factor will apply to that line. For short lines of only a few kilometres length (16.10) for the sphere will suffice, but for greater accuracy we have to use the appropriate equation from the list III.18, III.19, III.44 or III.45 derived for the spheroid. Moreover the solution for a line of finite length is to apply Simpson's Rule for integration which, applied to the line between the points (E_1, N_1) and (E_2, N_2) and which has the mid-point (E_m, N_m) which is defined by the mean of the coordinates of the ends. If the scale-factors calculated for the two terminal points are F_1 and F_2 respectively, and that for the mid-point is F_m, the scale-factor to be applied to the whole line is

$$1/F' = 1/6[1/F_1+4/F_m+1/F_2] \tag{III.48}$$

This is now applied to the chord distance between the points

$$K^2 = (E_2-E_1)^2+(N_2-N_1)^2 \tag{III.49}$$

and the required spheroidal distance is

$$s = K/F' \tag{III.50}$$

(t–T) Correction from E, N

This is normally determined for two points, such as *A* and *B*, whose grid coordinates are known. There are two solutions, first for the correction to be made at *A* for the line *AB* and secondly for the correction to be applied at *B* for the line *BA*

$$(t-T)_{AB} = (2E_A+E_B)(N_A-N_B)(1/6\rho\nu) \tag{III.51}$$

and

$$(t-T)_{BA} = (2E_B+E_A)(N_B-N_A)(1/6\rho\nu) \tag{III.52}$$

Azimuth from grid bearing

The grid bearing between two points is, of course,

$$\alpha_{AB} = \tan^{-1}[(E_1-E_2)/(N_1-N_2)] \tag{III.53}$$

It follows that the azimuth Z of a geodesic may be determined from this by application of both the arc-to-chord correction and the convergence:

$$Z_{AB} = \alpha_{AB} - (t - T) + \gamma \qquad \text{(III.54)}$$

Transverse Mercator by double-projection

The Biernacki-Rapp solution:

This was described on p. 350 where equations (16.35) and (16.36) describe the determination of spherical latitude and longitude from geodetic latitude and longitude

$$\Phi = P_0 + P_2\lambda^2 + P_4\lambda^4 \ldots \qquad \text{(III.55)}$$

$$\Lambda = P_1\lambda + P_3\lambda^3 + P_5\lambda^5 + \ldots \qquad \text{(III.56)}$$

The values of P_i may be derived from the following expansions in n:

$$P_0 = \varphi - [(3n/2) - (9n^3/16)] \sin 2\varphi + [(15n^2/16) + \ldots] \sin 4\varphi - [(35n^3/48) + \ldots] \sin 6\varphi \qquad \text{(III.57)}$$

$$P_1 = [1 + (n/2) + (n^2/4) - (7n^3/96) + \ldots] + [n/2 - (n^2/8)(5n^3/12) + \ldots] \cos 2\varphi - [(3n^2/8) + (5n^3/96) + \ldots] \cos 4\varphi + [(7n^3/24) + \ldots] \cos 6\varphi \qquad \text{(III.58)}$$

$$P_2 = [n/4 + (21n^2/32) + (25n^3/64) + \ldots] \sin 2\varphi + [n/8 + (3n/32) - (15n^3/12) + \ldots] \sin 4\varphi \qquad \text{(III.59)}$$

$$P_3 = [n/24 + 31n^2/96 + \ldots] + [n/6 + 37n^2/48 \ldots +] \cos 2\varphi + [n/8 + 15n^2/32 + \ldots] \cos 4\varphi + [n^2/48 \ldots] \cos 6\varphi \qquad \text{(III.60)}$$

$$P_4 = [(7n/96) + 9n^2/16 + \ldots] \sin 2\varphi + [(n/12) + (83n^2/128) + \ldots] \sin 4\varphi + [(n/32) + (13n^2/48) + \ldots] \sin 6\varphi \qquad \text{(III.61)}$$

$$P_5 = [n/240 + \ldots] + [17n/480 + \ldots] \cos 2\varphi + [n/16 \ldots] \cos 4\varphi + [n/32 + \ldots] \cos 6\varphi \qquad \text{(III.62)}$$

The remaining unknown is that of R, the radius of the conformal sphere. The derivation of this has been given in equations (16.37)–(16.39). Finally, having established values for Φ, Λ and R, the plane coordinates of a point may be determined from equations (16.45) and (16.46).

References

Adams, O.S. (1921): *Latitude Developments Connected with Geodesy and Cartography* . . . , USC & GS Spec. Publ. No. 67, 112 pp.

Admiralty (1954): *Admiralty Navigation Manual*, Vol. III, London, HMSO, 437 pp.

Admiralty (1960): *Admiralty Navigation Manual*, Vol. II, London, HMSO, 297 pp.

Admiralty (1965): *The Admiralty Manual of Hydrographic Surveying*, Vol. 1, London, Hydrographic Department, 671 pp.

Agajelu, S.I. (1987): On conformal representation of geodetic positions in Nigeria. *Surv. Rev.* **29**, 3–12.

AMS (1950): Universal Transverse Mercator Grid Tables for Latitudes 0°–80°. Vol. I: *Transformation of Coordinates from Geographic to Grid*; Vol. II: *Transformation of Coordinates from Grid to Geographic*; Vol. III: *Coordinate for 5-minute Intersections.* AMS 1, December 1950. Army Map Service, Corps of Engineers. Department of the Army, Washington, DC.

Anderson, E.W. (1968): *The Principles of Navigation*, London, Hollis and Carter, 635 pp.

Arden-Close, Sir Charles (1944): Correspondence; nomenclature of map projections. *Emp. Surv. Rev.* **54**, 345.

Arden-Close, Sir Charles (1952): A forgotten pseudo-zenithal projection. *Geogr. J.*, **118**, 237.

Arthur, D.W.G. (1978): Orthogonal transformations. *Am. Cartog.* **5**, 72–74.

Baker, J.G.P. (1986): The 'Dinomic' world map projection. *Cartogr. J.*, **23**, 66–67.

Barney, G.O. (ed.) (1980): *The Global 2000 Report to the President*, Vol. 1, Harmondsworth, Penguin Books, 766 pp.

Barton, B.A. (1976): A note on the transformation of spherical coordinates. *Am. Cartog.*, **3**, 161–168.

Benny, A.H. (1983): Automatic relocation of ground control points in Landsat imagery. *Int. J. Remote Sensing*, **4**, 335–342.

Beresford, P.C. (1953): Map projections used in polar regions. *J. Inst. Navigation*, **6**, 29–37.

Bernstein, R. (1976): Digital image processing of earth observation sensor data. *IBM J. Res. Devel.*, **20**, 40–57.

Bernstein, R. (ed.) (1983): *Image geometry and rectification*, Chap. 21 in Colwell (1983), Vol. 1, pp. 873–922.

Berrill, A.R. and Clerici, E. (1977): Statistical tests of digital rectification of LANDSAT image data. *Austral. Surv.*, **28**, 497–504.

Biernacki, F. (1949): *Teoria odwzorowan powierzchni dla geodetow i kartografow* (Theory of Representation of Surfaces for Surveyors and Cartographers). Warsaw, Glowny Urzqd Pomiarow Kraju, Prace Geodezynjego Instytuto Naukowo-Badawczego, no. 4. English translation, US Dept. of Commerce, 1965.

Billingsley, F.C. (ed.) (1983): Data processing and reprocessing. Chapter 17 in Colwell (1983), Vol. 1, pp. 719–792.

Blakemore, M. (ed). (1986): *Proceedings Auto Carto London*, Vol. 1: *Hardware, Data Capture and Management Techniques*; Vol. 2: *Digital Mapping and Spatial Information Systems.* Auto Carto London, Ltd, for International Cartographic Association.

Boggs, W.W. (1929): A new equal-area projection for world maps. *Geogr. J.*, **73**, 241–245.

Bomford, A.G. (1962): Transverse Mercator arc-to-chord and finite distance scale factor formulae. *Emp. Surv. Rev.* **16**(125), 318–327.

Bomford, G. (1962): *Geodesy*, 2nd edn., Oxford, Clarendon Press, 561 pp.
Bomford, G. (1980): *Geodesy*, 4th edn., Oxford, Clarendon Press, 855 pp.
Borgeson, W.T., Batson, R.M. and Kieffer, H.H. (1985): Geometric accuracy of Landsat-4 and Landsat-5 thematic mapper images. *Photogramm. Eng. Remote Sensing*, **51**, 1893–1898.
Born, G.H., Lame, D.B. and Mitchell, J. L. (1984): A survey of oceanographic satellite altimetric missions. *Mar. Geod.*, **8**, 3–16.
Bowring, B.R. (1976): Transformation from spatial to geographical coordinates. *Surv. Rev.* **23**(181), 323–327.
Bowring, B.R. (1989): Transverse Mercator equations obtained from a spherical basis. *Surv. Rev.*, **30**(233), 125–133.
Bowring, B.R. (1990a): New ideas about isometric latitude. *Surv. Rev.*, **30**(236), 270–280.
Bowring, B.R. (1990b): The Transverse Mercator projection – a solution by complex numbers. *Surv. Rev.* **30**(237), 325–342.
Brandenberger, A.J. and Gosh, S.K. (1985): The world's topographic and cadastral mapping operation. *Photogramm. Eng. Remote Sensing*, **51**, 437–444.
Brazier, H.H. (1951): A skew orthomorphic projection with particular reference to Malaya. Conf. Commonwealth Survey Officers, 1947, *Report of Proceedings*, London, HMSO, pp. 42–62.
Briesemeister, W. (1953): A new oblique equal-area projection. *Geogr. Rev.*, **43**, 260–261.
Briesemeister, W. (1959): A new equal area projection for the future. Second International Cartographic Conference, Chicago, 1958. *Proceedings*. Frankfurt, Verlag des Instituts für Angewandte Geodäsie, pp. 60–63.
Briggs, D. and Mounsey, H. (1989): Integrating land resource data into a European geographical information system: practicalities and problems. *Appl. Geogr.*, **9**, 5–20.
Bryant, N.A., Zobrist, A.L., Walker, R.E. and Gokhman, B. (1985): An analysis of Landsat thematic mapper P-product internal geometry and conformity to earth surface geometry. *Photogramm. Eng. Remote Sensing*, **51**, 1435–1447.
Bugaevskii, L.M. (1982): Kriterii otsenki pri v'bore kartograficheskikh proektsiy. *Geod. Aerofot.*, **3**, 92–96.
Bunge, W. (1962): *Theoretical Geography*. Lund Studies in Geography, Ser. C. General and Mathematical Geography No. 1. Lund, 2nd edn., 1966, 289 pp.
Burkard, R.K. (1964): *Geodesy for the Layman*, St Louis, Aeronautical Chart and Information Center (now DMAAC), 93 pp.
Burnside, C.D. (1985): *Mapping from Aerial Photographs*, 2nd ed., London, Collins, 348 pp.
Burrough, P.A. (1986): *Principles of Geographical Information Systems for Land Resource Assessment*, Oxford, Clarendon Press, 194 pp. Monograph on Soil and Resources Survey, No. 12.
Cahill, B.J.S. (1909): An account of a new land map of the world. *Scot. Geogr. Mag.*, **25**, 449–469.
Cahill, B.J.S. (1929): Projections for world maps. *Monthly Weather Rev.* **57**(4), 128–133.
Cahill, B.J.S. (1934): A world map to end world maps. *Geogr. Annal.*, **16**, 97–108.
Caley, A. (1843): On the motion of rotation of a solid body. *Camb. Math. J.*, **3**, 224–232.
Caley, A. (1846): On the rotation of a solid body round a fixed point. *Camb. Dublin Math. J.*, **1**, 167–173, 264–274.
Chorley, Lord Roger (1988): Some reflections on the handling of geographical information. *Int. J. GIS*, **2**, 3–9.
Chovitz, B.H. (1956): A general formula for ellipsoid-to-ellipsoid mapping. *Boll. di Geod. e Sci. Affini*, **15**, 1–20.
Chovitz, B.H. (1981): Modern geodetic earth reference models. *EOS, Trans. Am. Geophys. Union*, **62**, 65–67.
Clark, D. (ed. J. Clendinning) (1944): *Plane and Geodetic Surveying for Engineers*, Vol. 2: *Higher Surveying*, 3rd edn., revised and enlarged by J. Clendinning, London, Constable, 511 pp.
Clark, D. (ed. J.E. Jackson) (1973): *Plane and Geodetic Surveying for Engineers*, Vol. 2. *Higher Surveying*, 6th edn., revised and largely rewritten by J.E. Jackson, London, Constable, 292 pp.

Clarke, F.L. (1973): Zone to zone transformation on the Transverse Mercator Projection. *Austral. Surv.*, **25**, 293–302.
Clough-Smith, J.H. (1966): *An Introduction to Spherical Trigonometry*, Glasgow, Brown, Son and Ferguson, 111 pp.
Cocks, K.D., Walker, P.A. and Pavey, C.A. (1988): Evolution of a continental-scale geographical information system. *Int. J. GIS*, **2**(3), 263–280.
Colvocoresses, A.P. (1974): Space oblique Mercator. *Photogramm. Eng.*, **40**, 921–926.
Colwell, R.N. (ed.-in-chief) (1983): *Manual of Remote Sensing*, 2nd edn. in two vols. Falls Church, Va, American Society of Photogrammetry, 2145 pp.
Cooper, M.A.R. (1974): *Fundamentals of Survey Measurement and Analysis*, London, Crosby Lockwood Staples, 107 pp.
CSO (1947): Conference of Commonwealth Survey Officers, 1947. *Report of Proceedings*, London, HMSO, 1951.
CSO (1951): Conference of Commonwealth Survey Officers, 1951. *Report of Proceedings*. London, HMSO, 1956.
'Correspondent' (1945–46): Correspondence; nomenclature of map projections. *Emp. Surv. Rev.*, **60**(238), 64–85.
Cotter, C.H. (1966): *The Astronomical and Mathematical Foundations of Geography*, London, Hollis and Carter, 244 pp.
Cotter, C.H. (1969): *The Complete Nautical Astronomer*, London, Hollis and Carter, 336 pp.
Craster, J.E.E. (1929): Some equal-area projections of the sphere. *Geogr. J.*, **74**, 471–474.
Cross, P. (1986): Prospects for satellite and inertial positioning methods in land surveying. *Land and Minerals Surveying*, **4**, 196–203.
Curran, P.J. (1984): Geographic Information Systems. *Area*, **16**, 153–158.
Curran, P.J. (1985): *Principles of Remote Sensing*, London, Longman.
Curran, P.J. (1987): Remote sensing methodologies and geography. *Int. J. Remote Sensing*, **8**, 1255–1275.
Dahlberg, R.E. (1962): Evolution of interrupted map projections. *Int. Jahrb. Kartographie*, **2**, 36–54.
Dale, P.F. (1976): *Cadastral Surveys within the Commonwealth*. Overseas Research Publication No. 23, London, HMSO, 281 pp.
Davison, G.J. (1986): Ground control pointing and geometric transformation of satellite imagery. *Int. J. Remote Sensing*, **7**, 65–74.
Day, J.W.R. (1990): The finite distance scale factor formulae for Transverse Mercator, Documental Mercator and skew orthomorphic grids correct to fourth order terms. *Surv. Rev.* **30**(236), 244–258.
Depuydt, F. (1983): The equivalent quintuple projection. *Int. Jahrb. Kartographie*, **23**, 63–74.
DMS (1958): *Universal Transverse Mercator Grid Tables for the International, Clarke 1880, Clarke 1866, Bessel and Everest Spheroids*. Prepared by the Directorate of Military Survey, the War Office, April 1958.
Dowman, I.J. (1978): Topographic mapping from space photography: further developments. *Photogramm. Rec.*, **9**(52), 513–522.
Dowman, I.J. (1985): Images from space: the future for satellite photogrammetry. *Photogramm. Rec.*, **11**(65), 507–513.
Dowman, I.J. and Mohamed, M.A. (1981): Photogrammetric applications of Landsat MSS imagery. *Int. J. Remote Sensing*, **2**, 105–113.
Doytsher, Y. and Shmutter, B. (1981): Transformation of conformal projections for graphical purposes. *Canad. Surv.*, **35**, 395–404.
Driencourt, L. and Laborde, J. (1932): *Traité des projections des cartes géographiques à l'usage des cartographes et des geodesiens*, Paris, Hermann et cie, 4 vols.
Dyer, G.C. (1971): Polar navigation – a new Transverse Mercator technique. *J. Inst. Navigation*, **24**, 484–495.
Eckert, M. (1906): Neue Entwurfe fur Erdkarten. *Pet. Mitt.*, **52**, 97–109.
Eckhardt D. (1983): Cover illustration of a SEASAT geoid. EOS. *Trans. Am. Geophys. Union*, **64**, 25.
Edoga, A.C. (1981): Computation of the foot-point latitude in coordinate transformation. *Surv. Rev.* **26**(202), 192–194.

Edwards, A.T. (1953): *A New Map of the World: the Trystan Edwards Homolographic Projection*, London, Batsford, reprinted 1955, 18 pp.
Fawcett, C.B. (1949): A new net for a world map. *Geogr. J.*, **114**, 68–70.
Fiala, F. (1957): *Mathematische Kartographie*, Berlin, VEB Verlag Technik, 316 pp.
Field, N.J. (1980): Conversions between geographical and Transverse Mercator coordinates. *Surv. Rev.* **25**(195), 228–230.
Fisher, J. and Lockley, R. M. (1954): *Sea Birds*, London, Collins, 320 pp.
Fister, F.I. (1980): Some remarks on Landsat MSS pictures. 14th Congress of the International Society of Photogrammetry, Commission IV. *Int. Arch. Photogramm.*, pp. 214–222.
Freer, T.St.B. and Irwin, K.J. (1951): Proposals for a new navigation chart. *J. Inst. Navigation*, **4**, 66–80.
Frolov, Y.S. and Maling, D.H. (1969): The accuracy of area measurement by point counting techniques. *Cartogr. J.*, **6**, 21–35.
Fusco, L., Frei, U. and Hsu, A. (1985): Thematic Mapper: operational activities and sensor performance at ESA/Earthnet. *Photogramm. Eng. Remote Sensing*, **51**, 1299–1314.
Gardner, A.C. and Creelman, W.G. (1965): *Navigation*, Oxford, Pergamon Press, 251 pp.
Garland, G.D. (1965): *The Earth's Shape and Gravity*, Oxford, Pergamon Press, 183 pp.
Ginzburg, G.A. (1966): Novyy variant polikonicheskoy proektsii, *Geod. i Kartogr.*, **3**, 55–57.
Ginzburg, G.A. and Salmanova, T.D. (1957): *Atlas diya vybora kartograficheskikh proektsiy*. Trudy TsNIIGAiK, Vyp 110, Moscow, Geodezizdat, 240 pp.
Ginzburg, G.A. and Salmanova, T.D. (1962): *Primeneniye v matematicheskoy kartografii metodov chislennogo analiza*. Trudy TsNIIAiGK, Vyp 153, Moscow, Geodezizdat, 80 pp.
Ginzburg, G.A. and Salmanova, T.D. (1964): *Posobie po Matematicheskoy Kartografii*. Trudy TsNIIGAiK, Vyp 160, Moscow, Nedra, 456 pp.
Goode, J.P. (1925): The homolosine projection: new device for portraying the earth's surface entire. *Assoc. Am. Geogr. Annals*, **15**, 119–125.
Goodier, R. (ed.) (1971): *The Application of Aerial Photography to the Work of the Nature Conservancy*. Edinburgh, Nature Conservancy, 185 pp.
Goussinsky, B. (1951): On the classification of map projections. *Emp. Surv. Rev.*, **11**(80), 75–79.
Graff, D.R. (1988): Coordinate conversion from NAD 27 to NAD 83. *J. Surv. Eng.*, **114**, 125–130.
Gray, J. (1979): *Ideas of Space*, Oxford, Clarendon Press.
Green, N.P., Finch, S. and Wiggins, J. (1985): The 'state of the art' in geographical information systems. *Area*, **17**, 295–301.
Gringorten, I.I. (1972): A square equal-area map of the world. *J. Appl. Meteorol.*, **11**, 763–767.
Gugan, D.J. (1987): Practical aspects of topographic mapping from SPOT imagery. *Photogramm. Rec.*, **12**(69), 349–355.
Gugan, D.J. (1989): Future trends in photogrammetry. *Photogramm. Rec.*, **13**(73), 79–84.
Gugan, D.J. and Dowman, I.J. (1988): Accuracy and completeness of topographic mapping from SPOT imagery. *Photogramm. Rec.*, **12**(72), 787–796.
Hagger, A.J. (1950): Air navigation in high latitudes. *Polar Rec.*, **39**, 484–495.
Hammer, E. (1889): *Über die geographische wichtigsten Kartenprojektionen*, Stuttgart, J.B. Metzlerscher Verlag, 148 pp.
Harley, J.B. (1975): *Ordnance Survey Maps: a Descriptive Manual*. London, HMSO, 200 pp.
Heaney, G.F. (1967): Sir George Everest. *Geogr. J.*, **133**, 209–211.
Heiskanen, W.A. and Vening Meinesz, F.A. (1958): *The Earth and its Gravity Field*, New York, McGraw-Hill, 470 pp.
Hinks, A.R. (1912): *Map Projections*, Cambridge, Cambridge University Press, 126 pp.
Hinks, A.R. (1921a): *Map Projections* 2nd edn., Cambridge, Cambridge University Press, 158 pp.
Hinks, A.R. (1921b): On the projection adopted for the allied maps on the western front. *Geogr. J.*, **57**, 448–451.

Hinks, A.R. (1941): Murdoch's third projection. *Geogr. J.*, **97**, 358–362.
Hinks, A.R. (1942): Making the British Council map. *Geogr. J.*, **100**, 123–130.
Hinks, A.R. (1944): *Maps and Surveys*, 5th edn, Cambridge, Cambridge University Press, 311 pp.
Hirvonen, R.A. (1971): *Adjustment by Least Squares in Geodesy and Photogrammetry*, New York, Ungar, 261 pp.
Honick, K.R. (1967): Pictorial navigation displays. *Cartogr. J.*, **4**, 72–81.
Horn, B.K.P. and Woodham, R.J. (1979): Landsat MSS coordinate transformation. *1979 Machine Processing of Remotely Sensed Data Symposium* Purdue University, pp. 59–68.
Hotine, M. (1945–1946): Correspondence; nomenclature of map projections. *Emp. Surv. Rev.*, **53**, 276–278; **55**, 37; **61**, 276–277.
Hotine, M. (1946–7): The orthomorphic projection of the spheroid. *Emp. Surv. Rev.*, **8/9**, 62–66.
Hotine, M. (1947): Discussion, in CSO (1947).
Hotine, M. (1955): Discussion, in CSO (1951).
Hough, F.W. (1955): The Universal Transverse Mercator grid (with particular reference to Africa). Conference of Commonwealth Survey Officers 1951, *Report of Proceedings*, Paper no. 7, London, HMSO.
Hristow, V.K. (1943): *Die Gauss-Krüger'schen Koordinaten auf dem Ellipsoid.* Leipzig and Berlin (see also under Khristov).
ICA (International Cartographic Association) (1973): *Multilingual Dictionary of Technical Terms in Cartography*, Viesbaden, Franz Steiner Verlag, 573 pp.
ICA (International Cartographic Association) (1984–88): *Basic Cartography for Students and Technicians*, Vol. 1, International Cartographic Association, 1984, 206 pp.; Vol. 2 (ed. R.W. Anson), London, Elsevier Applied Science Publishers, 141 pp.
Jackson, J.E. (1978): Transverse Mercator projection. *Surv. Rev.* **24**(188), 278–285.
Jackson, J.E. (1980): *Sphere, Spheroid and Projections for Surveyors*, London, Granada, 138 pp.
Jackson, M.C. (1988): Personal communication.
Jackson, M.C. (1990): A map projection for the classroom. *Cartogr. J.*, **27**, 44–45.
Jackson, M.J. and Mason, D.C. (1986): The development of integrated geoinformation systems. *Int. J. Remote Sensing*, **7**, 723–740.
Jankowski, P. and Nyerges, T. (1989): Design considerations for MaPKBS-map projection knowledge-based system. *Am. Cartogr.*, **16**, 85–95.
Junkins, J.L. and Turner, D.A. (1978): A distortion-free map projection for analysis of satellite imagery. *J. Astronaut. Sci.*, **26**(3), 211–234.
Kadmon, N. (1975): Data-bank derived hyperbolic scale equitemporal town maps. *Int. Jahrb. Kartographie*, **15**, 47–54.
Kadmon, N. (1983): Photographic, polyfocal and polar-diagrammatic mapping of atmospheric pollution. *Cartogr. J.*, **20**, 121–126.
Kadmon, N. and Shlomi, E. (1978): A polyfocal projection for statistical surfaces. *Cartogr. J.*, **15**, 36–41.
Kanazawa, K. (1984): Techniques of map drawing and lettering. In: ICA, *Basic Cartography for Students and Technicians*, Vol. 1, International Cartographic Association, pp. 112–180.
Keuning, J. (1955): The history of geographical projections until 1600. *Imago Mundi*, **12**, 1–24.
Khristov, V.K. (1957): *Koordinaty Gaussa-Krugera na ellipsoide vrashcheniya.* Moscow, Geodezizdat, 263 pp.
King-Hele, D.G. (1960): *Satellites and Scientific Research*, London, Routledge and Kegan Paul, 180 pp.
King-Hele, D.G. (1964): The shape of the earth, *J. Inst. Nav.*, **17**, 1–16.
King-Hele, D.G. (1967): The shape of the earth, *Sci. Am.*, **217**, 67–76.
King-Hele, D.G. (1976): The shape of the earth, *Science*, **192**(4246), 1293–1300.
King-Hele, D.G. (1978): The gravity field of the Earth. *Phil. Trans. R. Soc. Lond.*, **294A**, 317–328.
Konecny, G. (1976): Mathematical models and procedures for the geometric restitution of

remote sensing imagery. XIII Congress of the International Society for Photogrammetry, Helsinki, Finland, July, *Int. Arch. Photogramm*, **XXI**, Part 3.
Konecny, G. (1979): Methods and possibilities for digital differential rectification. *Photogram. Eng. Remote Sensing*, **45**, 727–734.
Krarup, T. (1969): *A Contribution to the Mathematical Foundation of Physical Geodesy*, Publication No. 4, Danish Geodetic Institute, Copenhagen, 80 pp.
Kratky, V. (1974): Cartographic accuracy of ERTS. *Photogramm. Eng.*, 203–212.
Kratky, V. (1975): Image transformation in satellite mapping. Conference of Commonwealth Survey Officers, 1975, *Report of Proceedings*, Paper K4, London, Ministry of Overseas Development, Vol. II.
Kratky, V. (1981): Spectral analysis of interpolation. *Photogrammetria*, **37**, 61–72.
Lame, D.B. and Born, G.H. (1982): SEASAT Measurement Evaluation: Achievements and Limitations. *J. Geophys. Res.*, **87**(C5), 3175–3178.
Lauf, G.B. and Young, F. (1961): Conformal transformations from one map projection to another using divided differences interpolation. *Bull. Geod.*, **61**, 191–212.
Lee, L.P. (1944): The nomenclature and classification of map projections. *Emp. Surv. Rev.*, **7**, 190–200.
Lee, L.P. (1945): The Transverse Mercator projection of the spheroid. *Emp. Surv. Rev.* **8**(58), 142–152.
Lee, L.P. (1946): The nomenclature of map projections. *Emp. Surv. Rev.*, **8**(60), 217–219.
Lee, L.P. (1965): Some conformal projections based on elliptic functions. *Geogr. Rev.*, **55**, 563–580.
Lee, L.P. (1973): The conformal tetrahedric projection with some practical applications. *Cartogr. J.*, **10**, 22–28.
Lee, L.P. (1974): A conformal projection for a map of the Pacific. *New Zealand Geogr.*, **30**, 75–77.
Lee, L.P. (1976): *Conformal Projections based on Elliptic Functions*, Cartographica, Monograph No. 16.
Lenox-Conyngham, Sir Gerald (1944): Correspondence; nomenclature of map projections. *Emp. Surv. Rev.*, **53**, 276 (see also under following authors: Sir Charles Arden-Close, M. Hotine, L.P. Lee, and 'Correspondent' (1944–46).
Lucas, E.F. (1977): Transformation from local to UTM coordinates. *Surv. Rev.*, **24**(183), 42–48.
McCaw, G.T. (1940): The Transverse Mercator projection; a critical examination. *Emp. Surv. Rev.*, **5**(35), 285–296.
Macdonald, R. R. (1968): An optimum continental projection. *Cartogr. J.*, **5**, 46–57.
McGrath, G. M. (1976): From Hills to Hotine. *Cartogr. J.*, **13**, 7–21.
Mackinder, Halford (1902): *Britain and the British Seas*, London, Oxford University Press, 375 pp.
Maguire, D.J., Goodchild, M.F. and Rhind, D.W. (eds) (1991): *Geographical Information Systems; Overview, Principles and Applications*, Longmans Scientific and Technical UK (In press).
Makris, G.A. (1982): Foot-point latitude: correspondence arising from the paper by A.C. Edoga. *Surv. Rev.*, **26**(205), 345–347.
Maling, D.H. (1960): A review of some Russian map projections. *Emp. Surv. Rev.*, **15**, 203–215, 255–266, 294–303.
Maling, D.H. (1962): The Hammer–Aitoff projection and some modifications. *Proc. Cartogr. Symp. Edinburgh 1962*. Geography Dept., University of Glasgow, pp. 41–57.
Maling, D.H. (1968a): How long is a piece of string? *Cartogr. J.*, **5**, 147–156.
Maling, D.H. (1968b): The terminology of map projections. *Int. Jahrb. Kartogr.*, **8**, 11–65.
Maling, D.H. (1971): Photogrammetric techniques relevant to the Nature Conservancy's use of air photography. In Goodier (1971), pp. 132–149.
Maling, D.H. (1983): 'A little, round, fat, oily man of God'; Rev. Patrick Murdoch and his contributions to eighteenth century cartography and geography. *Cartogr. J.*, **20**, 110–118.
Maling, D.H. (1984): Mathematical cartography. In: *ICA Basic Cartography for Students and Technicians*, Vol. 1, International Cartographic Association, pp. 32–78.

Maling, D.H. (1989): *Measurements from Maps*, Oxford, Pergamon Press, 547 pp.
Maling, D.H. (1991): Chapter 10, Coordinate systems and projections. In Maguire *et al.* (1991).
Mark, D.M. and Lauzon, J.O. (1985): Approaches for quadtree-based geographical information systems at continental or global scales. AUTO CARTO VII, *Proceedings*, Washington, pp. 355–364.
Marsh, J.G. and Martin, T.V. (1982a): The SEASAT altimeter mean sea-surface model. *J. Geophys. Res.*, **85**, 3269–3280.
Marsh, J.G., Martin, T.V. and McCarthy, J.J. (1982b): Global mean sea-surface computation using GEOS-3 altimeter data. *J. Geophys. Res.*, **87**, 10955-10964.
Mather, P.M. (1987): *Computer Processing of Remotely-sensed Images*, Chichester, Wiley, 352 pp.
Maurer, H. (1935): *Ebene Kugelbilder*, Pet. Mitt. Ergänzungsheft No. 221, Gotha, 87 pp.
Meade, B.K. (1987): Program for computing Universal Transverse Mercator (UTM) coordinates for latitudes north or south and longitudes east and west. *Surv. Mapping*, **47**, 37–49.
Meneguette, A.A.C. (1985): Evaluation of Metric Camera photography for mapping and co-ordinate determination. *Photogramm. Rec.*, **11**(66), 699–709.
Methley, B.D.F. (1986): *Computational Models in Surveying and Photogrammetry*, Glasgow, Blackie, 346 pp.
Mikhail, E.M. (1976): *Observations and Least Squares*, New York, IEP-Dun-Donnelly Harper & Row, 497 pp.
Mikhail, E.M. and Gracie, G. (1981): *Analysis and adjustment of survey measurements.* New York, Van Nostrand Reinhold, 340 pp.
Miller, O.M. (1941): A conformal map projection for the Americas. *Geogr. Rev.*, **31**, 100–104.
Miller, O.M. (1953): A new conformal projection for Europe and Asia [*sic.*; should read Africa]. *Geogr. Rev.*, **43**, 405–409.
Ministry of Defence (1962): *Manual of Military Engineering*, Vol. XIII: *Survey*, Part XII: *Cartography*, London, Ministry of Defence, 323 pp.
Ministry of Defence (1966): *Manual of Military Engineering*, Vol. XIII: *Survey*, Part VI: *Survey Computations*, 1st edn, London, Ministry of Defence, 243 pp.
Ministry of Defence (1973): *Manual of Map Reading*, London, Ministry of Defence, HMSO, 142 pp.
Ministry of Defence (1978): *Manual of Military Engineering*, Vol. XIII: *Survey*, Part VI: *Survey Computations*, 2nd edn, London, Ministry of Defence, 243 pp.
Moritz, H. (1980a): Geodetic reference system 1980. *Bull Géod.*, **54**, 395–405.
Moritz, H. (1980b): *Advanced Physical Geodesy*, Karlsruhe, Herbert Wichmann Verlag, and Tunbridge Wells, Abacus Press, 500 pp.
Morrison, J.E. (1989): The revolution in cartography in the 1980's. In: Rhind and Taylor (1989), pp. 169–185.
Mounsey, H. and Tomlinson, R.F. (eds): *Building Databases for Global Science*, London, Taylor and Francis, pp. 129–137.
NRSC (1985): *United Kingdom, National Remote Sensing Centre, Data Users Guide.* Annually, but 1985 version used here.
Nyerges, T.L. and Jankowski, P. (1989): A knowledge base for map projection selection. *Am. Cartogr.*, **16**, 29–38.
Olliver, J.G. (1981): A zone to zone transformation method for the Universal Transverse Mercator grid (Clarke 1880 spheroid). *Surv. Rev.*, **26**(199), 36–45.
Ordnance Survey (1950): *Constants, Formulae and Methods used in Transverse Mercator Projection*, London, HMSO, 32 pp.
Ordnance Survey (1951): *The Projection for Ordnance Survey Maps and Plans and the National Reference System*, London, HMSO, 4 pp.
Ordnance Survey (1967): *The Retriangulation of Great Britain*, London, HMSO, 2 vols.
Pavlov, A.A. (1967): Preobrazovanie kartogrficheskikh proektsii na elektronnykh vychislitel'nych mashinakh. *Geod. Aerofot.*, **4**, 27–33.
Peake, E.R.L. (1947): The activities of ICAO with particular reference to those of its map division. Conference of Commonwealth Survey Officers, 1947, *Report of Proceedings*, London, HMSO, 1951, pp. 181–202.

Proctor, D.W. (1986): The capture of survey data. *Autocarto London*, **1**, 227–236.
Rapp, R.H. (1975): *Lecture Notes on Map Projection.* Dept. of Geodetic Science, Ohio State University.
Rauhala, U.A. (1972): New solutions for fundamental calibration and triangulation problems. *Svensk Lantmäteritidskrift*, **4**(2).
Redfearn, J.C.B. (1948): Transverse Mercator formulae. *Emp. Surv. Rev.*, **9**(69), 318–322.
Reignier, F. (1957): *Les systemes de projection et leurs applications a la Geographie, a la navigation, a la topometrie etc.*, Paris, Institut Geographique National, 2 vols.
Rhind, D.W. and Taylor, D.R.F. (eds) (1989): *Cartography, Past, Present and Future*, London, Elsevier Applied Science Publishers on behalf of the International Cartographic Association, 193 pp.
Richards, J.A. (1986): *Remote Sensing Digital Image Analysis*, Berlin, Springer-Verlag, 281 pp.
Richardus, P. and Adler, R. K. (1972): *Map Projections, for Geodesists, Cartographers and Geographers*, Amsterdam, North-Holland, 174 pp.
Robinson, A.H. (1952): *The Look of Maps*, Madison, University of Wisconsin Press, 105 pp.
Robinson, A.H. (1953): Interrupting a map projection: a partial analysis of its value. *Assoc. Am. Geogr. Annals*, **43**, 216–225.
Rouleau, B. *et al.* (1984): Theory of cartographic expression and design. In: ICA, *Basic Cartography for Students and Technicians*, Vol. 1, International Cartographic Association, pp. 81–111.
RGS (Royal Geographical Society) (1989): What a world of difference! *1 Kensington Gore, the Newsletter of the Royal Geographical Society*, November.
Royal Society (1966): *Glossary of Technical Terms in Cartography*, London, Royal Society, 84 pp.
Ruffhead, A. (1987): An introduction to least-squares collocation. *Surv. Rev.* **29**(224), 85–94.
Sear, W.J. (1967): The projection story. *Cartography*, **6**, 64–72.
Seymour, W.A. (ed). (1980): *A History of the Ordnance Survey*, Folkestone, Dawson, 394 pp.
Shmutter, B. (1981): Transforming conic conformal to TM coordinates. *Surv. Rev.*, **26**(201), 130–136.
Snyder, J.P. (1977): A comparison of pseudocylindrical map projections. *Am. Cartogr.*, **4**, 59–81.
Snyder, J.P. (1978): The space oblique Mercator projection. *Photogramm. Eng. Remote Sensing*, **44**, 585–596.
Snyder, J.P. (1981): *Space Oblique Mercator Projection – Mathematical Development.* US Geological Survey, Bulletin 1518, Washington, Government Printing Office, 108 pp.
Snyder, J.P. (1982a): Geometry of a mapping satellite. *Photogramm. Eng. Remote Sensing*, **48**, 1593–1602.
Snyder, J.P. (1982b): *Map Projections used by the U.S. Geological Survey.* US Geological Survey, Bulletin 1532, Washington, Government Printing Office, 313 pp.
Snyder, J.P. (1984): A low-error conformal map projection for the 50 States. *Am. Cartogr.*, **11**, 27–39.
Snyder, J.P. (1985): *Computer-assisted Map Projection Research.* US Geological Survey Bulletin 1629, Washington, Government Printing Office, 157 pp.
Snyder, J.P. (1987a): *Map projections – A Working Manual.* US Geological Survey Professional Paper 1395, Washington, Government Printing Office, 383 pp.
Snyder, J.P. (1987b): Differences due to projection for the same USGS quadrangle. *Surv. Mapping*, **4**, 199–206.
Snyder, J.P. (1987c): Labeling projections on published maps. *Am. Cartogr.*, **14**, 21–27.
Snyder, J.P. (1987d): 'Magnifying-glass' azimuthal map projections. *Am. Cartogr.*, **14**, 61–68.
Snyder, J.P. and Steward, H. (1988): *Bibliography of Map Projections.* US Geological Survey Bulletin 1856, Washington, Government Printing Office, 110 pp.
Snyder, J.P. and Voxland, P. M. (1989): *An Album of Map Projections.* US Geological Survey Professional Paper 1453, Washington, Government Printing Office, 249 pp.

Sodano, E.M. (1965): General non-iterative solution of the inverse and direct geodetic problems. *Bull. Geod.*, **75**, 103–109.
Soler, T. and Hothem, L.D. (1988): Coordinate systems used in geodesy: basic definitions and concepts. *J. Surv. Eng.*, **114**, 84–97.
Spiess, E. and Brandenberger, C. (1986): Transformation of digital and analogue map data by interpolation methods – an important tool in small-scale map compilation. *Int. Jahrb. Kartogr.*, **26**, 149–181.
Sprinsky, W.H. (1987): Transformation of positional geographic data from paper-based map products. *Am. Cartogr.*, **14**, 359–366.
Stanley, H.R. *et al.* (1979): Scientific results of the GEOS-3 mission. *J. Geophys. Res.*, **84**(B8), 3779–4079.
Steiner, D. and Kirby, M.E. (1976): Geometrical referencing of Landsat images by affine transformation and overlaying of map data. *Photogrammetria*, **33**, 41–75.
Stigant, G.B. (1947): Discussion, in CSO (1947).
Strasser, G. (1975): The toise, the yard and the metre – the struggle for a universal unit of length. *Surv. Mapping*, **35**, 25–46.
Thomas, P. D. (1952): *Conformal Projections in Geodesy and Cartography*, US Coast and Geodetic Survey, Spec. Publ. No. 251, 142 pp.
Thompson, E.H. (1969): *An Introduction to the Algebra of Matrices with some Applications*, London, Adam Hilger, 229 pp.
Thompson, E.H. (1973): Review of 'Coordinate Systems and Map Projections'. *Photogramm. Rec.*, **7**, 755–758.
Thompson, E.H. (1975): A note on conformal map projections. *Surv. Rev.*, **23**(175), 17–28.
Tobler, W.R. (1963): Cartographic area and map projections. *Geogr. Rev.*, **53**, 59–78.
Tobler, W.R. (1964a): *Geographical Coordinate Computations*, Part I: *General Considerations*, University of Michigan, Department of Geography, Technical Report No. 2, ONR Task No. 389–137.
Tobler, W.R: (1964b): *Geographical Coordinate Computations*, Part II: *Finite Map Projection Distortions*, University of Michigan, Department of Geography, Technical Report No. 2, ONR Task No. 389–137.
Tobler, W.R. (1988): Resolution, resampling, and all that. In: Mounsey and Tomlinson (1988), pp. 129–137.
Tobler, W. R. (1989): Spherical quadrilateral to map scale conversion. *Am. Cartog.*, **16**, 54.
Tobler, W.R. and Zi-tan Chen (1986): A quadtree for global information storage. *Geogr. Anal.*, **18**, 360–371.
Tomlinson, R. (1988): The impact of the transition from analogue to digital cartographic representation. *Am. Cartogr.*, **15**, 249–261.
UN (1970): The status of world topographic mapping. *World Cartogr.*, **10**, 1–96.
UN (1976): The status of world topographic mapping. *World Cartogr.*, **14**, 1–70.
UN (1979): The status of world topographic mapping. *World Cartogr.*, **17**, 3–115.
US Army (1955): *A Guide to the Compilation and Revision of Maps*, Dept of the Army Technical Manual TM 5-240. Washington, Dept. of the Army, 167 pp.
US Hydrographic Office (1932): *Useful Tables from the American Practical Navigator*, Washington, US Government Printing Office, 220 pp.
Vakhrameeva, L.A. and Bugaevsky, U.L. (1985): Issledovanie peremenno-masshabnykh proektsii dlya sotsial'no-ekonomicheskikh kart. (Research into variable scale projections). *Geodez. Aerofot.*, **3**, 95–99.
Van Wie, P. and Stein, M. (1977): A Landsat digital image rectification system. *IEEE Trans.*, **GE-15**, 130–137.
Vincenty, T. (1971): The meridional distance problem for desk computers. *Surv. Rev.*, **121**(161), 136–140.
Vincenty, T. (1985): Correspondence about the paper 'Non-iterative solution of the φ-equation'. *Surv. Mapping*, **45**, 265–266.
Vincenty, T. (1987): Conformal transformations between dissimilar plane coordinate sysems. *Surv. Mapping*, **47**, 271–274.
Vincenty, T. (1988): A note on approximate transformations of coordinates. *Surv. Mapping*, **48**, 207–211.

Vincenty, T. (1989): The flat earth concept in local surveys. *Surv. Mapping*, **49**, 101–102.
Wade, E.B. (1986): Impact of North American datum of 1983. *J. Surv. Eng.*, **112**, 49–62.
Wagner, K.-H. (1949): *Kartographische Netzentwürfe*, Leipzig, Bibliographisches Institut, 263 pp.
Wagner, K.-H. (1982): Bemerkungen zum Umbeziffern von Kartennetzen. *Kartogr. Nachr.*, **6**, 211–218.
Watts, D. (1970a): Some new map projections of the world. *Cartogr. J.*, **7**, 41–46.
Watts, D. (1970b): Pseudo-conical world maps. *Cartogr. J.*, **7**, 101–102.
Welch, R. and Usery, E.L. (1984): Cartographic accuracy of Landsat-4 MSS and TM image data. *IEEE Trans.*, **GE-22**, 281–288.
Welch, R., Jordan, T.R. and Ehlers, M. (1985): Comparative evaluations of the geodetic accuracy and cartographic potential of Landsat-4 and Landsat-5 Thematic Mapper image data. *Photogramm. Eng. Remote Sensing*, **51**, 1249.
Wiggins, J.C., Hartley, W.P., Higgins, M.J. and Whittaker, R.J. (1986): Computing aspects for a large geographic information system for the European Community, *Autocarto London*, **2**, 28–43.
Williams, J. M. (1979): *Geometric Correction of Satellite Imagery*, London, HMSO, 20 pp, 10 figs, Royal Aircraft Establishment Technical Report No. 79121. Departmental reference: Space 569.
Williams, W.B.P. (1982): The Transverse Mercator projection – simple but accurate formulae for small computers. *Surv. Rev.*, **25**, 307–320.
Winterbottom, H.S.L. (1934): *The National Plans*. Ordnance Survey Professional Papers, New Series No. 16, London, HMSO, 100 pp.
Wray, T. (1974): *The Seven Aspects of a General Map Projection*. Cartographica Monograph No. 11, 72 pp.
Wray, T. and Weiss, C.C. (1982): Auxiliary spheres on common spheroids. *Canad. Surv.*, **36**, 191–196.
Wright, J.W. (1988): The plain surveyor's guide to geographic information systems. *Land Minerals Surv.*, **6**, 400–409.
Wu, Zhong-xing and Yang, Qi-he (1981): A research on the transformation of map projections in computer-aided cartography: Paper presented at the 10th International Cartographic Conference, Tokyo, unpublished manuscript. Chinese language version published in *Acta Geodet Cartogr. Sinica*, **10**, 20–44.
Young, A.E. (1920): *Some Investigations in the Theory of Map Projections*, London, Royal Geographical Society Technical Series No. 1, 76 pp.
Young, A.E. (1923): Note to Professor J.D. Everett's application of Murdoch's third projection. *Geogr. J.*, **62**, 359–361.
Zakatov, P.S. (1962): *A Course in Higher Geodesy*. Translated from the Russian and published for the National Science Foundation, Washington, DC, by the Israel Program for Scientific Translation, Jerusalem.
Zienkiewicz, O.C. (1977): *The Finite Element Method*, Maidenhead, 3rd Edition, McGraw-Hill, 787 pp.

Index of Names

Where an article or book has more than one author, only the surname of the first appears in this index. All joint authors are listed in the REFERENCES pp 450–459.

Index of Map Projections

Subject Index