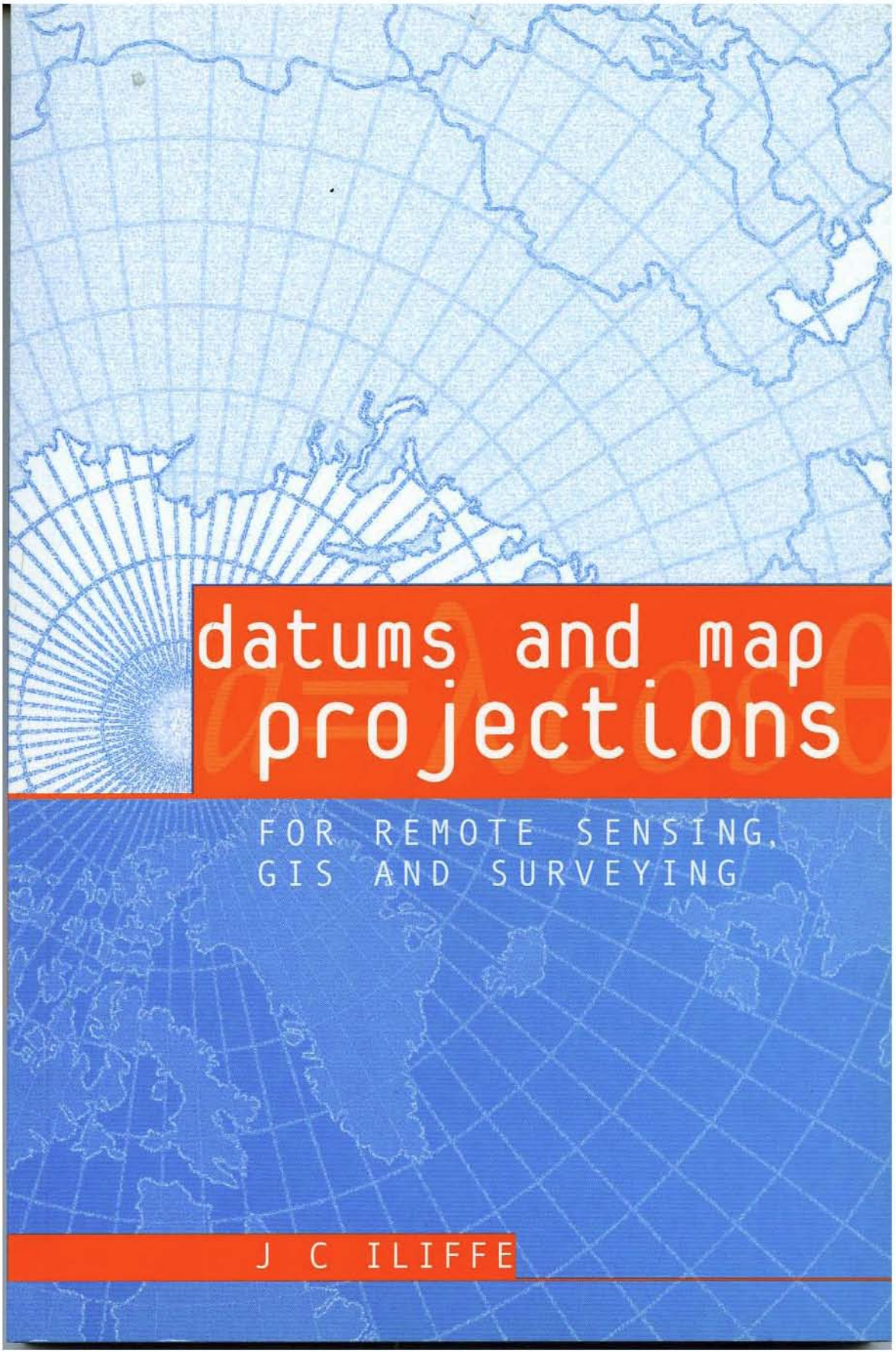
datums and map projections
FOR REMOTE SENSING, GIS AND SURVEYING
J C ILIFFE

DATUMS AND MAP PROJECTIONS

Datums and Map Projections

for remote sensing, GIS, and surveying

Jonathan Iliffe
Department of Geomatic Engineering
University College London

Whittles Publishing

CRC Press
Boca Raton London New York Washington, D.C.

Typeset by
Whittles Publishing Services

Published by
Whittles Publishing,
Roseleigh House,
Latheronwheel,
Caithness, KW5 6DW,
Scotland, UK

Distributed in North America by
CRC Press LLC,
2000 Corporate Boulevard N.W.,
Boca Raton,
FL 33431, USA

ISBN 1-870325-28-1
USA ISBN 0-8493-0884-4

Printed by Bell & Bain Ltd., Glasgow

Contents

Preface and acknowledgements		**vii**
1	**Introduction**	**1**
	1.1 Background	1
	1.2 Coordinates and datums	2
2	**Two- and three-dimensional coordinate systems**	**8**
	2.1 Introduction	8
	2.2 Spherical coordinates	8
	2.3 Spheroidal coordinates	9
	2.4 Cartesian coordinates	12
3	**Height and the geoid**	**14**
	3.1 The geoid	14
	3.2 Reference surfaces for height	19
4	**Global, regional, and local datums**	**19**
	4.1 Global datums	23
	4.2 Local and regional datums	26
5	**The global positioning system**	**33**
	5.1 Introduction	33
	5.2 System overview	33
	5.3 Positioning with codes	35
	5.4 Differential GPS using codes	38
	5.5 GPS phase measurements	41
6	**Aspects of datum transformations**	**45**
	6.1 Introduction	45
	6.2 Knowledge of separation, N	45
	6.3 Knowledge of height, H	45
	6.4 Knowledge of datum transformation parameters	47
	6.5 Datum transformations for precise applications	50
7	**Fundamentals of map projections**	**58**
	7.1 Introduction	58
	7.2 Spheres and spheroids	59

7.3 Grids and graticules 59
7.4 Scale factor 60
7.5 Developable surfaces 61
7.6 Preserved features 63
7.7 Computational aspects 65
7.8 Designing a projection 66

8 Cylindrical projections 68
8.1 Cylindrical equidistant projection 68
8.2 Cylindrical equal area projection 70
8.3 Mercator projection 72
8.4 Transverse Mercator projection 74
8.5 Oblique Mercator projection 77

9 Azimuthal projections 79
9.1 General azimuthal projection 79
9.2 Azimuthal equidistant projection 79
9.3 Azimuthal equal area projection 81
9.4 Azimuthal stereographic (conformal) projection 82
9.5 Gnomonic projection 83

10 Conic projections 84
10.1 General conic projection 84
10.2 Conic equidistant projection 87
10.3 Albers (conic) equal area projection 88
10.4 Lambert conformal conic projection 88

11 Summary of information required 92
11.1 Formulae 92
11.2 Parameters 92

12 Direct transformations 95
12.1 Compatibility of coordinate systems 95
12.2 Ground control 97
12.3 Plane transformations 98
12.4 Unknown projections: measuring from maps 101

13 Case studies 105
13.1 Transformation of GPS data into a local datum 105
13.2 A projection for Australia 109
13.3 Establishment of maritime boundaries on a projection 111
13.4 Two-dimensional transformation of satellite imagery 113
13.5 Two-dimensional transformation of GPS data 115
13.6 Determining the parameters of an unknown projection 116

Appendix 123
A1 Spherical coordinates 123
A2 Basic geometry of the ellipsoid 123
A3 Determination of transformation parameters by least squares 127

References 142
Index 144

Preface

This book has been written as a practical guide to the problems that are commonly encountered when using datums and map projections. It is aimed at students and practitioners in the areas of remote sensing, geographic information systems, and surveying.

Several trends in recent years have been responsible for bringing the importance of datums and projections to the attention of a wider community of users. One has been the development of new methods of acquiring spatial data, such as the global positioning system and satellite remote sensing. Another has been the introduction of geographic information systems for handling and manipulating data in digital form. Taken together, these have brought about a situation where users are no longer reliant on paper maps provided by a single source, but can acquire and customise spatial data as their needs dictate. To assure the quality of the products derived from these processes, it is essential that the diverse coordinate frameworks upon which the data are based should be fully understood.

This book is for users with different levels of background knowledge of the subject, and therefore provides some fundamental definitions as well as more detailed explanations. The emphasis throughout is placed upon the solution of practical problems, giving alternatives where standard procedures may not be adequate, and giving worked examples and appendices with useful formulae.

Those dipping into the book to find answers to specific queries on transforming data will find a "route diagram" of different transformations and procedures in Chapter 1, which gives a reference to where more complete descriptions and formulae can be found. Three chapters then follow on different types of coordinate systems and datums: Chapter 2 considers two- and three-dimensional representations of Earth coordinates; Chapter 3 looks at vertical datums; and Chapter 4 explores the ways in which datums are established on a global, regional, or local basis.

Chapter 5 gives an introduction to the global positioning system, and explores the problems that can arise with datums when using this. Chapters 7 to 1 1 introduce the fundamentals of map projections, and look at the different types in some detail. Models and procedures for transforming directly between data sets, based on the identification of common points, are treated in Chapter 12.

Chapter 13 introduces several practical projects in the form of case studies, which together illustrate the range of problems that may be encountered when working with data sets on different datums or when attempting to carry out computations in projected coordinate systems.

Finally, I should like to place on record my gratitude to many colleagues in the Department of Geomatic Engineering at University College London for the assistance that they have given me in writing this book. To Paul Cross, Ian Dowman, Arthur Allan, John Arthur, David Chapman, and Joel Barnes, my thanks are due for many interesting discussions on different aspects of the text and the application of techniques in different fields.

I should also like to thank Ryan Keenan, Joel Barnes, and Joanne Gilman for their assistance with many of the diagrams.

That said, all remaining errors in the book are mine.

JONATHAN ILIFFE

1
Introduction

1.1 Background

This book is designed as a practical guide for those working with spatially referenced data to the problems that may be encountered with datums and map projections. It is aimed at those working in surveying, remote sensing, geographic information systems, and related areas, and therefore covers a wide range of scales and accuracy targets. People encountering these topics may have very different starting points in terms of their level of knowledge: those who are aware that they need to know something about the subject, but are not yet sure what, would do well to commence by reading section 1.2 before proceeding further.

Until recently, an in-depth knowledge of datums was generally confined to a fairly small group of scientists and to geodesists working in government survey departments. Most surveying was carried out on local scales using terrestrial equipment (such as theodolites, levels and distance measurers) to establish positions with respect to nearby control points: although the map projection was usually a consideration, the question of the datum was of minimal importance. For the general user requiring access to spatial data (on land use, for example), almost the only practical source was a published map.

Two developments in recent years have been principally responsible for changing this state of affairs. One has been the explosion in the use of spatial data that has been brought about by the development of geographic information systems (GIS) for handling and manipulating data in digital form. The other has been the development of techniques such as the global positioning system (GPS) and satellite (or airborne) remote sensing, which have made available entirely new methods of acquiring accurate data. Moreover, these are techniques that, due to their global, space-based nature, have broken free completely of the localised survey. In short, this is a classic case of supply and demand: more and more data is available from a variety of different sources, and more and more people have the means and the need to make use of it.

Supply, demand and – potentially – confusion. A situation now exists where it is quite common for the means of acquiring data to be using a reference system that is completely different from the one in which the data will ultimately be required. This is a particular case of the problem of combining data sets, in which new data is to

be put alongside archive material (which could have been referenced to one of many scores of possible datums and an uncountable number of possible map projections), but the problem is a more general one. A few examples will serve to illustrate this.

- Geo-referencing a satellite image with ground control points that have been established using GPS, and with others that have been obtained from a published map.
- Combining digital map data from two different survey organisations, for example as part of a cross-border collaboration between neighbouring states.
- Carrying out a survey with high precision GPS, and bringing it into sympathy with existing mapping in a local coordinate system.
- Navigating a ship or an aircraft using satellite systems, on charts that have been prepared in a local datum. Even a leisure user of handheld GPS receivers will need to make an appropriate correction when using them in conjunction with a map.

These examples are significant, in that they can each be said to represent an increasing trend. Indeed, it could be argued that the same forces that drive globalisation in the political and economic spheres are having their effect on trends in the area of spatial information. After all, a hydrographic surveyor in the English Channel and a land surveyor in Ghana could both simultaneously be using the same positioning satellite. They are probably not, at present, expressing the final result in the same datum, but all major international organisations with an interest in spatial data are involved in developing regional and global datums.

The examples given above focus on the problems of combining data from different sources. Another reason for understanding the nature of a reference system is that it is often necessary to carry out computations using the data: it must, after all, have been acquired for some purpose. Computations that use geographic or geodetic coordinates (latitude, longitude and height) require a particular branch of mathematics. Projections introduce certain distortions of distance.

In some situations, this is hardly a problem. For example, if Ordnance Survey digital data is being used to determine the optimum route on the road network between two points, the distortions of the projection are insignificant at the accuracy level required. To take an example at the other extreme of accuracy requirements, however, the establishment of maritime boundaries between states involves determining distances between median lines and coastal points. This computation can be done using geodetic coordinates, but not without difficulty. Using a projection is easier, but the distances will be distorted and the computation will be inexact: although this may be insignificant over small areas if an appropriate projection is used, it is not likely to be sufficiently accurate for large areas such as the North Sea. Where the limit lies, and what an appropriate projection is, are questions it is necessary to be able to answer when tackling problems of this nature.

The first intention of this book is to clarify the definitions of geodetic datums and map projections, and to explore some of their important characteristics. The procedures for transforming between data sets under a variety of different circumstances will then be discussed.

This is information which, in theory at least, is available in most textbooks on geodesy. In practice, however, there are two problems that can cause difficulties. Firstly, datums and transformations usually form a limited part of the content of geodetic textbooks, and the relevant information can be difficult to extract and is rarely aimed at the non-specialist who still has a real need to interpret geodetic data. Secondly, and more importantly, the treatment in most standard texts tends to assume that all necessary information is available and stops short of explaining what to do when it is not. Anyone with any practical experience of manipulating spatial data will know that not infrequently some apparently essential information is missing: nothing is known about the geoid, for example, or the exact datum and projection of a digitised map are unknown.

The second principal aim of this book is therefore to act as a practical guide to these problems and to show where particular procedures are necessary and where they are not, or to show what can be done about situations in which not all the information is available. This is reinforced by a series of case studies which examine particular problems or combinations of problems, and by an appendix that groups together some of the essential mathematical formulae and procedures for computations within a coordinate system or for transforming between them. In general, however, the assumption will be that most readers are well supplied with software packages to carry out the computations: interpreting what they are doing is the key issue.

1.2 Coordinates and datums

Consider each of the following statements:

- The height of the point is 3.122 m.
- The height above mean sea level is 10.983 m.
- The latitude is $32^\circ\ 10'\ 12.23''$.
- The northings of the point are 152 345.834.

All of these express coordinates with a great deal of precision: the other thing that they have in common is that they are all – to a greater or lesser extent – ambiguous statements, as they contain no information on the coordinate framework in which they are measured. What would render them unambiguous would be a clear definition of the datum that is associated with each one.

What is a datum? Many readers will have encountered the use of the expression with reference to heights: for example, in the statement 'this point is 3.2 m above the datum'. This therefore seems a reasonable place to start.

If, for example, an engineer is interested only in the relative heights of points within a construction project, and not in their relationship to the outside world, then it will be acceptable to designate one point as having an arbitrary height – 100 m for example – and finding the heights of all others with respect to this. Such an act has effectively defined a datum, because it has established the position of the origin of the coordinate system. It would also have been possible to define the datum by designating the height of any other point, in which case all heights would be different

(see Fig. 1.1). Thus, the definition of the datum is seen to be essential information that accompanies the given coordinates.

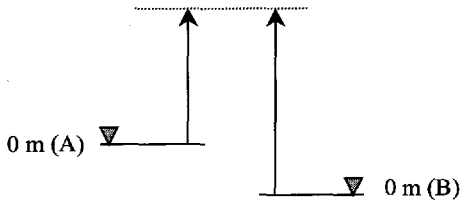

Figure 1.1 A point with different heights in datum A and datum B.

This is also the situation with two- or three-dimensional coordinate systems such as eastings and northings, or XYZ, or latitude and longitude. However, an additional complication is the fact that – without changing the datum – it is possible to express the coordinates in a different way. A useful example of this is the two-dimensional coordinate system shown in Fig. 1.2. In this example, the coordinates of the point P may be quoted in either the rectangular form (X,Y) or the polar form (r,θ): the point is that changing from one to the other is a relatively straightforward procedure that does not involve changing the point of origin, and does not imply a change of datum.

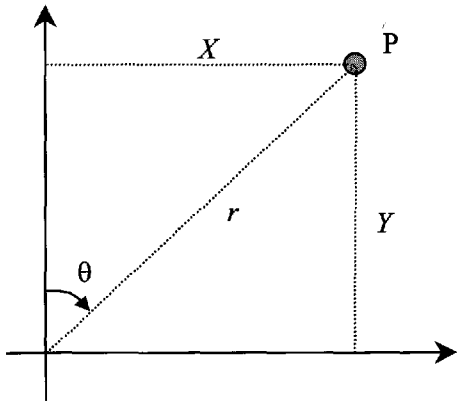

Figure 1.2 Rectangular and polar coordinates in one datum.

A similar situation exists in geodesy, but not all of the means of expressing position (or of converting between different forms) are as simple as in the above example. With this in mind, it is convenient to have a diagram of the different types of coordinate systems, with cross-references to where in this book fuller definitions or explanations of formulae may be found. This diagram can then be used as a 'route map' in navigating between different sets of coordinates.

In Fig. 1.3, each horizontal row represents a different datum – in this case labelled datums A and B. In this book, datums are covered in Chapter 4: international and satellite datums such as WGS-84 and ITRF are treated in section 4.1, and locally defined datums that are used as the basis of coordinates in individual countries or regions are discussed in section 4.2.

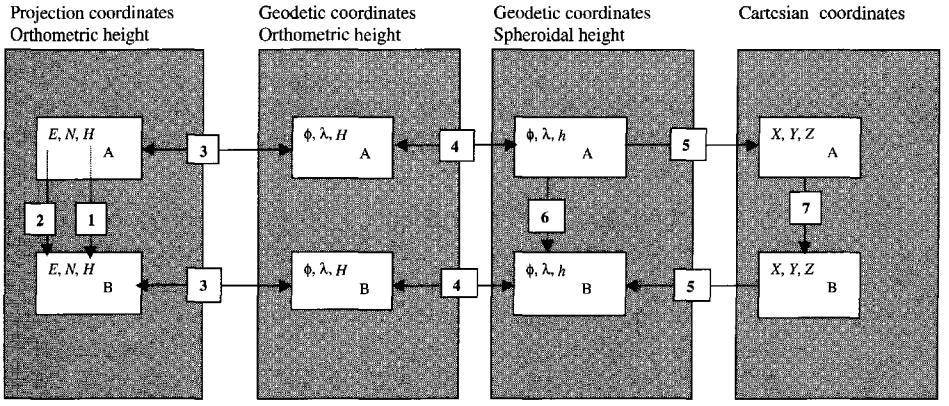

Figure 1.3 Complete procedure for transformations between different datums and projections.

Within each row, each vertically arranged box represents a different method of expressing coordinates. Brief, and not necessarily rigorous, definitions of these are as follows:

- *Projection coordinates:* An arrangement whereby the curved surface of the Earth is represented as a plane, thereby introducing distortions. Within this plane, a simple set of XY or east and north axes is defined. This topic is introduced in Chapter 7, and individual types of projection are discussed in Chapters 8–11. An example of the issues raised by carrying out computations within a projected coordinate system is given in section 13.3.

- *Orthometric heights (or heights above the geoid):* Heights defined above the irregular surface, the geoid, that closely approximates mean sea level. The geoid and its characteristics are explained in section 3.1, and the practicalities of establishing a vertical datum, and the consequences of the differences between the geoid and mean sea level, are discussed in section 3.2.

- *Geodetic coordinates:* Latitude and longitude defined with respect to a spheroid. The fundamental definitions of these are given in section 2.3.

- *Spheroidal (or ellipsoidal) heights:* Heights defined above a spheroid which has been established as the datum for a particular country or region, or on a global basis.

- *Cartesian coordinates:* Three-dimensional coordinates defined with respect to a set of axes with their origin at the centre of the spheroid used in the datum. These are discussed in section 2.4.

Any set of coordinates will be in one of the forms shown in Fig. 1.3. Most users of spatial data will want to do one of two things with the data:

- Carry out computations within the coordinate system that the data are in – for example to compute distances or angles between sets of points. In this case

reference should be made to the sections of this book that describe the nature and characteristics of the different types of coordinate systems.

- Convert the data to another coordinate form or another datum. This will involve one or more steps in the transformation process, which are labelled on the diagram and described below.

    1. The vertical datums used for orthometric heights are not actually related to the two-dimensional datums, and so H is on a separate datum to the eastings and northings (E,N). Section 3.2 discusses vertical datums, and points out that a change in the vertical datum changes the orthometric height by at most 1 m: for many purposes it will not be necessary to introduce a change, and so the orthometric height in the two systems will be the same.
    2. Direct conversion from one map to another is possible for low levels of accuracy, provided that common points can be identified in both systems. Chapter 12 explores the limitations of this approach, with examples being given in the case studies in sections 13.4 and 13.5. The mathematical basis for determining the parameters of the transformation is discussed in the appendix.
    3. The formulae for the conversion from projection coordinates to geodetic coordinates (and vice versa) depends on the type of projection. Chapter 11 summarises the information that will be required; section 12.4 discusses the options available if some of the parameters are not available, with the case study in section 13.6 giving an example of this.
    4. The conversion from geoidal or orthometric heights to spheroidal heights (and vice versa) requires a knowledge of the separation between the geoid and the spheroid, as discussed in section 3.1. The basic equations to be used are (3.1) and (3.2), which are given in section 3.2. Sections 6.2 and 6.5 respectively discuss the approaches that may be adopted for low and high accuracy applications if no information is available on the geoid–spheroid separation. It should be remembered that the value of the separation between geoid and spheroid is dependent on the datum that is being used.
    5. The conversion from geodetic to cartesian coordinates is straightforward, and requires only a knowledge of the parameters of the spheroid used in the datum. The formulae are given in section 2.4.
    6. Molodensky's formulae (section 4.2) can convert directly from geodetic coordinates on one datum to geodetic coordinates on another, given the change in spheroidal parameters and the offset between the two datums. This approach is easier for computations on a hand calculator, but otherwise the intermediate conversion to cartesian coordinates is conceptually easier (and is in any case a step that is simply hidden in the direct formulae). This book will generally adopt this approach.
    7. This step represents the actual datum transformation – a process that involves at least a three-dimensional shift, and possibly rotations and a scale

change. Section 4.2 introduces the topic; some of the problems that may be encountered are covered in Chapter 6. A full treatment of the determination of transformation parameters by least squares is given in the appendix, and the case study in section 13.1 gives a numerical example of the transformation of GPS data into a local datum.

2
Two- and three-dimensional coordinate systems

2.1 Introduction

As explained in section 1.2, a datum is a reference system in which the coordinates of spatial data may be expressed. It may be a three-dimensional system in cartesian or curvilinear form; it may be two-dimensional in a map projection or a locally defined system; or at its simplest it may be a one-dimensional system for expressing heights.

This chapter will consider the geometrical models that are used to define reference systems for expressing two- and three-dimensional coordinates of points.

In order to follow the definitions of the systems used, it is first necessary to consider the shape and size of the Earth.

2.2 Spherical coordinates

The first approximation that can be made to the shape and size of the Earth is that it is a sphere of radius 6371 km. Three-dimensional spherical coordinates can then be defined with respect to this shape (see Fig. 2.1):

- *latitude:* the angle north or south from the equatorial plane, ϕ
- *longitude:* the angle east or west from the Greenwich meridian, λ
- *height:* a distance in metres above (or below) the sphere, h.

The definition of latitude is a natural one, in that the poles are clearly defined points on the surface of the Earth, and the equator is the circle that bisects the two. On the other hand, the definition of longitude is to some extent arbitrary, as there are no physical reasons for choosing a particular meridian as the reference value. Historically, different prime meridians were used by different states, but the convention of using Greenwich is now universal.

The relationship between the definition of longitude in a particular datum and the establishment of a time datum should be remarked upon. In effect, the determination

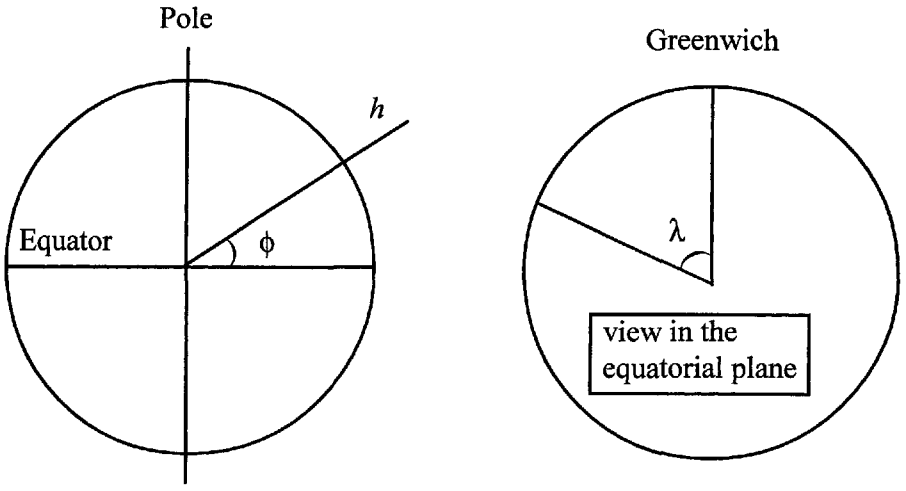

Figure 2.1 Spherical coordinates. ϕ, latitude; λ, longitude; h, height.

of the longitude with respect to the Greenwich meridian could only be carried out with a precision equivalent to the establishment of Greenwich Mean Time at the site of the datum. A 1 s second error in the determination of time would translate to a 15 arc second rotation of the datum with respect to the Greenwich meridian.

In conjunction with this coordinate system, it is useful to define the following terms:

- *parallels of latitude:* lines of equal latitude on the surface of the sphere
- *meridians:* lines of equal longitude.

Figure 2.2 shows the pattern that results from parallels and meridians depicted at intervals of 5°.

For many applications that do not require the highest accuracy, the sphere is an adequate representation of the Earth. Appendix section A1 summarises some of the formulae that can be used in conjunction with the spherical model, such as finding the distance and azimuth between points of known coordinates.

2.3 Spheroidal coordinates

A better approximation to the shape of the Earth is that it is an *ellipsoid of revolution*, often more conveniently called a *spheroid*. This surface is formed by an ellipse which has been rotated about its shortest (minor) axis, or by 'squashing' a sphere at the poles. The term *ellipsoid* can also be used for this shape and, although some would differentiate between an ellipsoid and a spheroid, in this context the two can be regarded as synonymous.

A spheroid formed in this way is referred to as an *oblate spheroid*, as opposed to a *prolate spheroid* which would be formed by extending the distance between the poles.

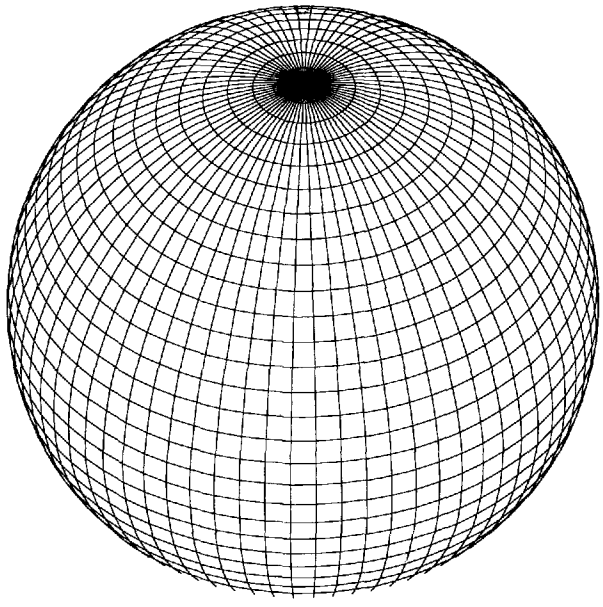

Figure 2.2 Meridians and parallels at 5° intervals.

As all the spheroids referred to in this text are oblate, the term spheroid can be used without ambiguity.

A spheroid is defined by the size of two parameters (see Fig. 2.3):

- the *semi-major axis*, a
- the *semi-minor axis*, b.

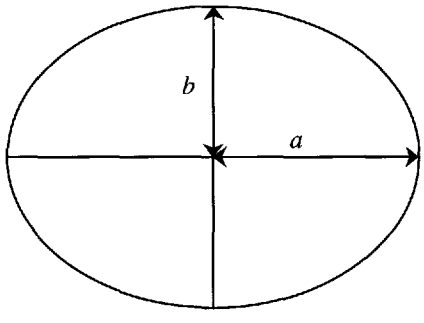

Figure 2.3 Defining parameters of a spheroid.

From these two parameters, it is possible to derive further definitions. Thus:

- *flattening*, f, is defined as

$$f = \frac{a-b}{a} \tag{2.1}$$

- *eccentricity*, e, is defined as

$$e^2 = \frac{a^2 - b^2}{a^2} \tag{2.2}$$

Furthermore, e, f and b can be related to each other as follows:

$$e^2 = 2f - f^2 \tag{2.3}$$

$$\sqrt{1 - e^2} = (1 - f) = \frac{b}{a} \tag{2.4}$$

Thus, a spheroid can be completely defined using the two parameters a and b, or a and f, or a and e, and the remaining parameters can be found as necessary.

A typical value for a would be 6 378 137 m and for f would be 1/298. A spheroid with this degree of flattening would, to the eye, appear indistinguishable from a sphere, and it must therefore be emphasised that most diagrams of the spheroid exaggerate the flattening for purposes of clarity.

It is now possible to define a set of coordinates with respect to this spheroid. Once again these are defined as *latitude*, *longitude* and *height*. These are shown in Fig. 2.4, from which it is apparent that the latitude and longitude are defined with respect to the direction of the *spheroidal normal*, a line from the point in question that is perpendicular to the spheroid.

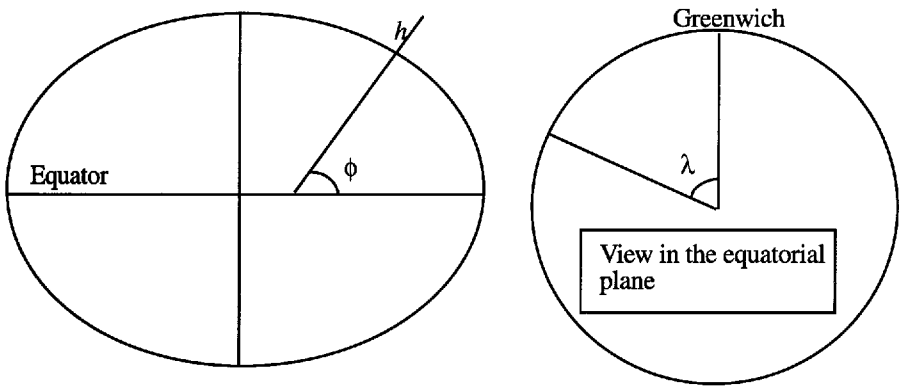

Figure 2.4 Geodetic coordinates. ϕ, latitude; λ, longitude; h, height.

Coordinates defined in this way are known as *geodetic coordinates*, and are the basis for all mapping systems. The distinction between geographic coordinates implying a spherical Earth model and geodetic coordinates implying a spheroidal one is useful, but by no means universally adopted: many texts use *geographic coordinates* as a generic term that also encompasses latitude and longitude defined with respect to a spheroidal model.

It is possible to carry out computations in a spheroidal coordinate system, particularly in the region close to its surface. In traditional land surveying, these would have

been used when computing coordinates in large scale geodetic surveys, a process that has largely been superseded by the use of cartesian coordinates (section 2.4) and the GPS (Chapter 5). Over shorter distances it has usually been possible to use a projected coordinate system. One of the few remaining areas where computations are needed in geodetic coordinates is in the definition of boundaries between states, or between mineral concessions. For this reason, a summary of some of the most useful formulae and procedures is given in Appendix section A2. Otherwise, any textbook on geodesy, such as Bomford (1980) or Torge (1991), gives further and more detailed information on spheroidal geometry.

2.4 Cartesian coordinates

The formulae involved in computations based on geodetic coordinates are complicated, and entirely inappropriate when considering observations made to satellites.

More appropriately, a set of cartesian coordinates (X, Y, Z) is defined with its origin at the centre of the spheroid. The Z axis is aligned with the minor axis of the spheroid (the 'polar' axis); the X axis is in the equatorial plane and aligned with the Greenwich meridian; the Y axis forms a right-handed system (see Fig. 2.5).

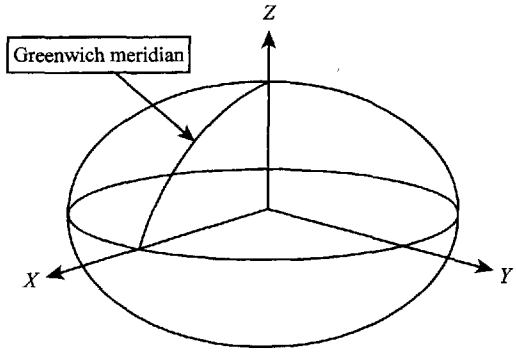

Figure 2.5 Cartesian coordinates.

Geodetic coordinates may be transformed to cartesian coordinates by a set of formulae which require a knowledge of the parameters of the spheroid:

$$\begin{aligned} X &= (\nu + h)\cos\phi\cos\lambda \\ Y &= (\nu + h)\cos\phi\sin\lambda \\ Z &= \{(1 - e^2)\nu + h\}\sin\phi \end{aligned} \tag{2.5}$$

where

$$\nu = \frac{a}{(1 - e^2\sin^2\phi)^{1/2}} \tag{2.6}$$

ϕ is the latitude, positive north; λ is the longitude, positive east; and h is the spheroidal height (the height above the spheroid). The reverse computation, in which geodetic

coordinates are found from cartesian ones, is also possible:

$$\tan\lambda = \frac{Y}{X} \tag{2.7}$$

$$\tan\phi = \frac{Z + \varepsilon b \sin^3 u}{p - e^2 a \cos^3 u} \tag{2.8}$$

$$h = (X^2 + Y^2)^{1/2} \sec\phi - \nu \tag{2.9}$$

where

$$p = (X^2 + Y^2)^{1/2} \tag{2.10}$$

$$\tan u = \frac{Z}{p}\frac{a}{b} \tag{2.11}$$

$$\varepsilon = \frac{e^2}{1 - e^2} \tag{2.12}$$

and all other terms are as defined above.

3
Height and the geoid

3.1 The geoid

The spheroid is a good approximation to the shape of the Earth, but not an exact representation of it. The shape of the Earth is given by the form of a surface which is everywhere perpendicular to the direction of gravity (such a figure is termed an *equipotential surface*). The force and direction of gravity are affected by irregularities in the density of the Earth's crust and mantle. It therefore follows that the form of an equipotential surface is somewhat irregular.

Of course, the fact that the spheroid is only an approximation of the shape of the Earth does not invalidate it as a model to be adopted in expressing coordinates. It is important to note that using geodetic coordinates based on a spheroid does not lead to errors when expressing the position of a point: it is simply a question of using a surface with a convenient mathematical expression onto which positions are projected.

To a very good approximation, the form of the *mean sea level* surface is equipotential, since the sea would be expected to be perpendicular to gravity. (In fact the seas and oceans contain permanent currents which cause a permanent slope with respect to the direction of gravity.) The true shape of the Earth is known as the *geoid*, and this can now be defined as *that equipotential surface that most closely corresponds to mean sea level.*

Worldwide, the difference between the geoid and mean sea level is at the most around 1 m (and changes slowly over wavelengths of tens of kilometres), and so for many purposes these two can be considered synonymous. Sometimes, however, it is important to note the difference.

The spheroid is a very good approximation to the geoid, but there are significant differences. If a spheroid is formed which best fits the geoid, the differences between the two amount to ± 100 m, with a global root mean square of around 30 m. Figure 3.1 shows the form of the geoid worldwide. The height of the geoid above the spheroid is known as the *geoid–spheroid separation*, or often just the *separation*, and is usually given the symbol N. This may be a positive or a negative quantity, as shown in Fig. 3.2.

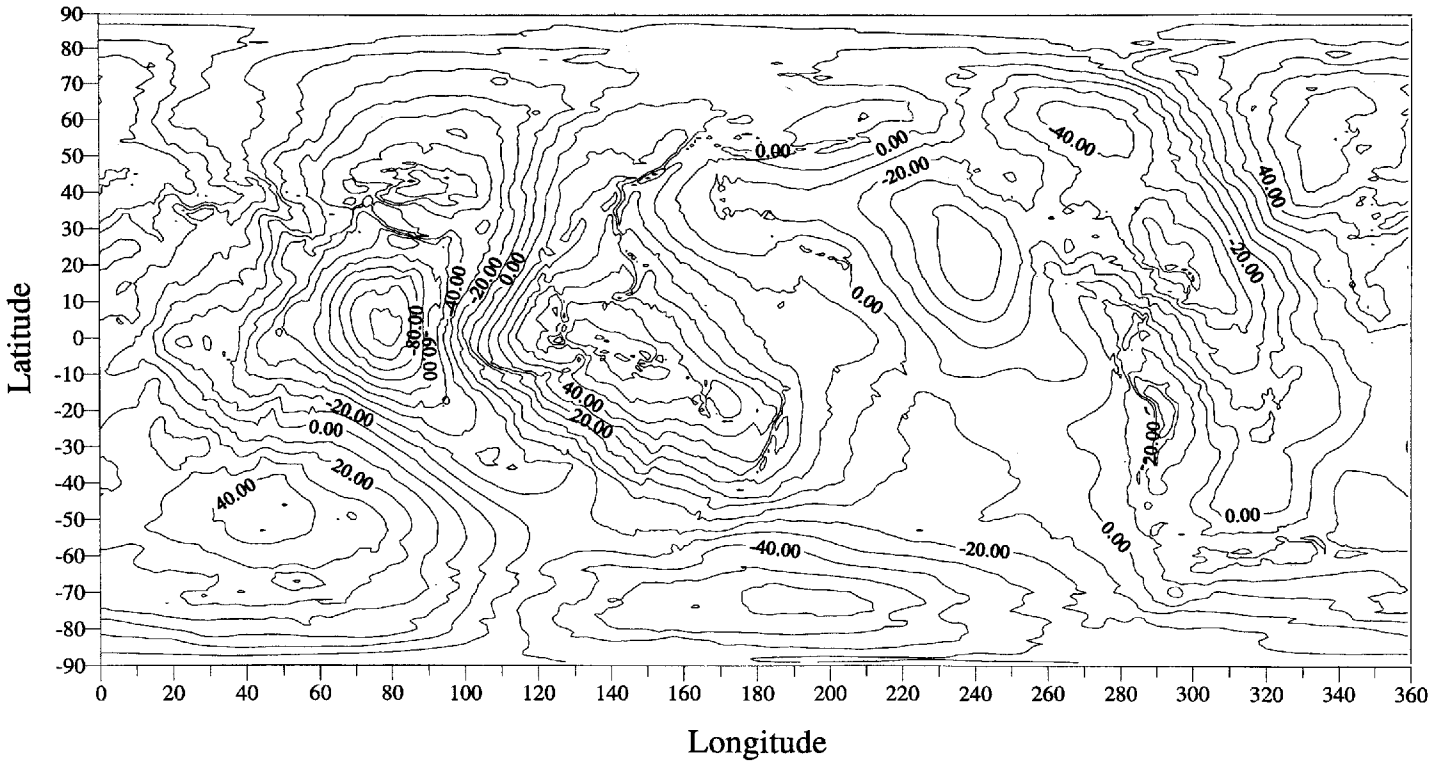

Figure 3.1 The geoid with respect to a best-fitting global datum. The values used are from the EGM96 model (NASA, 1996).

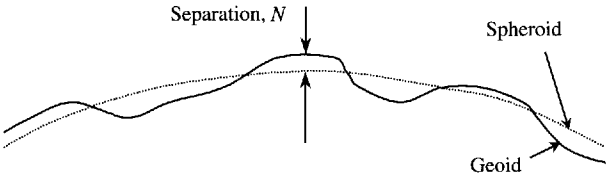

Figure 3.2 Sectional view of a spheroid and the geoid.

Some texts refer not to the height of the geoid above the spheroid, but to the *undulation* of the geoid. This is a rather unsatisfactory term, as the derivation of the word implies that the phenomenon under consideration is a wave, rather than a height. In this book N is called the separation, and the expression *undulations of the geoid* is reserved to refer to the presence of 'waves' in the part of the geoid in question. (A geoid without undulations, for example, would imply that at the accuracy required it appears to be at a constant height above the spheroid, or can be modelled simply as an inclined plane.)

Note also in Fig. 3.2 that the direction of the vertical (perpendicular to the geoid) is not usually coincident with a spheroidal normal. The angle between the two is termed the *deviation of the vertical*, and is usually resolved into a north–south component, ξ, and an east–west component, η. These angles typically amount to a few seconds of arc.

Astronomical coordinates may be defined with respect to the direction of the vertical in the same way that geodetic coordinates are defined with respect to the direction of the normal. These coordinates would be used in situations where a knowledge of the direction of the vertical is required (for example in determining the slope of the geoid), but they are somewhat irregular, and the geodetic coordinates remain the basis of the mapping system.

It must be emphasised that any diagram of the geoid will have to greatly exaggerate its deviation from the spheroid in order to be comprehensible. If it is recalled that the deviation between the two is less than 100 m over a body with dimensions of around 6400 km, then it will be appreciated that on a correct diagram the two surfaces would be indistinguishable. In fact, the geoid never actually 'doubles back on itself' – it is always convex – and therefore astronomical coordinates are always unique (it is impossible for two points to have the same latitude, a point not apparent from Fig. 3.2).

In many situations it is necessary to have a knowledge of the form of the geoid, either globally or for a specific region. This can be derived (with a great deal of effort) from gravity observations, observations of satellite orbits, satellite altimetry over the oceans, and other data sources. In developed countries with dense gravity observations the separation may be known to an accuracy of 5–10 cm. In other parts of the world it may be much worse than this. The point of reference would be the national mapping organisation for the country concerned or a university research department.

Globally, the geoid may be found from a set of coefficients of a *spherical harmonic expansion*, often referred to as a *global Earth model*. This expresses the geoid in terms of a series of functions over the sphere that are of increasingly small wavelength. The smallest wavelength that the model can express is a function of its highest *degree*: a model that is described as being complete to degree N_{MAX} can model wavelengths

down to $180°/N_{MAX}$. A global Earth model used to compute satellite orbits, for example, might be complete to degree 36, which would be sufficient to describe the long wavelength part of the geoid, but could not express any effects with a wavelength of less than 5°, or around 500 km.

Until recently, some of the most finely detailed Earth models were produced by the Ohio State University, and the model referred to as OSU91 (Rapp and Pavlis, 1990) was widely circulated and is still very much used. Being complete to degree 360, this models the form of the geoid down to wavelengths of around 0.5°, or 50 km, and is accurate to about 1–2 m. In some parts of the world the accuracy is rather worse than this, however, particularly in those areas where a dense network of gravity observations is not available.

OSU91 has now been superseded by a model that has been jointly determined by the National Imaging and Mapping Authority (NIMA) and the National Aeronautics and Space Administration (NASA), and is referred to as the Earth Geopotential Model 1996, or EGM96 (Rapp and Nerem, 1994) (see Fig. 3.2). This too is complete to degree 360, but gains an increased accuracy through the use of additional data, in particular gravity observations that were previously unavailable and new data from satellite missions.

The coefficients of the spherical harmonic expansion of EGM96, together with a computer program for determining the geoid from them, are freely available on the World Wide Web (NASA, 1996). Also available is a file of geoid point values determined on a 0.25° grid: a considerable amount of processing is saved if the geoid is interpolated from these values.

Clearly, then, a model such as EGM96 will solve the geoid problem in most parts of the world for users who require an accuracy of no better than 1 m. For accuracies greater than this, it is instructive to consider the size and characteristics of the part of the geoid that is not modelled by EGM96.

As stated at the start of this section, the irregular nature of the geoid is caused by the underlying irregularities in the Earth's gravity field. For longer wavelengths (which are usually of greater amplitude) this is usually due to deep-seated effects such as the interaction between the Earth's crust and mantle at the boundaries of tectonic plates (Lambeck, 1988). These features are usually adequately described by the global models. The shortest wavelength effects are mostly due to the irregularities of the terrain (Forsberg and Tscherning, 1981). Importantly, then, the size of the undulations of the geoid that exist over and above those effects modelled by EGM96 or OSU91 are very much correlated with the roughness of the terrain.

To get a feel for the characteristics of the geoid, examples will be shown of a recent project to determine the geoid in Zimbabwe, with data drawn from Nyamangunda (1997). Figure 3.3 shows a sample cross-section of the geoid at latitude 16° S, with the longitude ranging from 29° to 31° E: this is a distance of around 200 km. The smoother of the two lines in Fig. 3.3 is the underlying trend of the geoid that has been derived from the global model. The more complex line is a better approximation of the geoid, as this has been derived by using additional terrain and gravity data, and is therefore capable of picking up shorter wavelength effects.

On looking at a 2° cross-section of the geoid, the smoothing effect of the global model becomes apparent. It is clear, nevertheless, that the very large changes in the

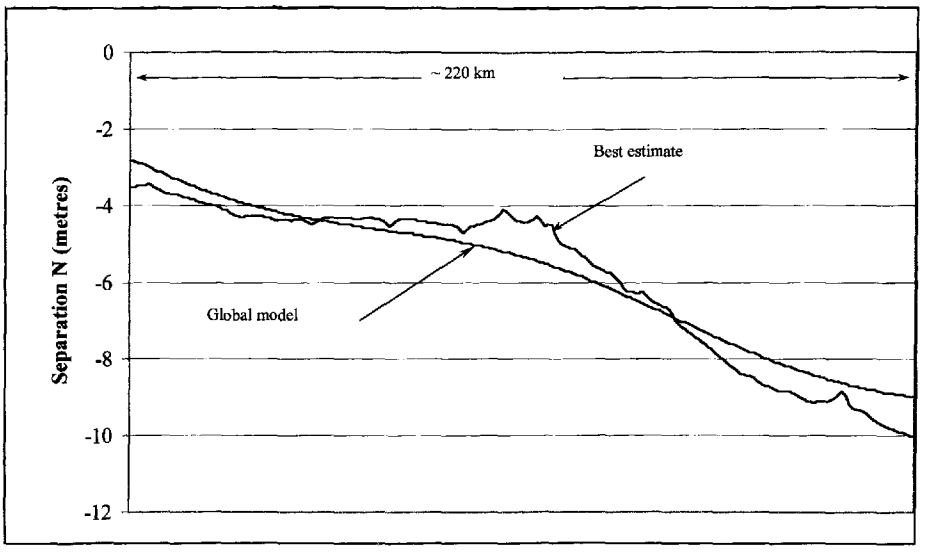

Figure 3.3 Cross-section of the geoid at latitude 16° S.

separation are generally quite well modelled by the global expansion. Taking the less smooth line to represent the actual geoid (although in fact the scarcity of data in this part of the world means that it does still have errors of a decimetre or so) some quite substantial undulations are still seen over distances of a few kilometres.

With the aid of Fig. 3.3, and referring also to the global picture of the geoid seen in Fig. 3.2, we now consider two practical questions about the geoid.

3.1.1 To what extent can the geoid–spheroid separation be considered a constant value?

Only to a very limited extent. If a project requires centimetric accuracy, it would not be safe to assume a constant value for the separation over distances of more than about 200 m.

For less precise applications, requiring an accuracy of, say, 2 m, Fig. 3.3 suggests that the separation might be considered constant over distances up to around 30–50 km. Referring to Fig. 3.1, however, it is notable that in some parts of the world the separation may change by 10 m over distances as short as 50 km.

To a certain extent, the situation is eased when using a local reference system, as described in section 4.2, as opposed to a global datum such as WGS84. A local reference system will usually have been optimised to fit the shape of the geoid in a particular region, and the slope of the geoid is usually less severe. In the UK, for example, the total range of the geoid separation is 0–5 m. But note that this is true only when expressing the geoid height with respect to the British datum: it is not true when expressing it with respect to WGS84. It should also be pointed out that this applies principally to slopes over longer distances: short-wavelength undulations are present whatever reference surface is used.

3.1.2 To what extent can the geoid be modelled as a uniform slope?

This question is of particular importance in precise GPS applications, as section 6.5 shows that a section of the geoid that approximates a sloping plane can effectively be ignored when using similarity transformations to tie into local control points.

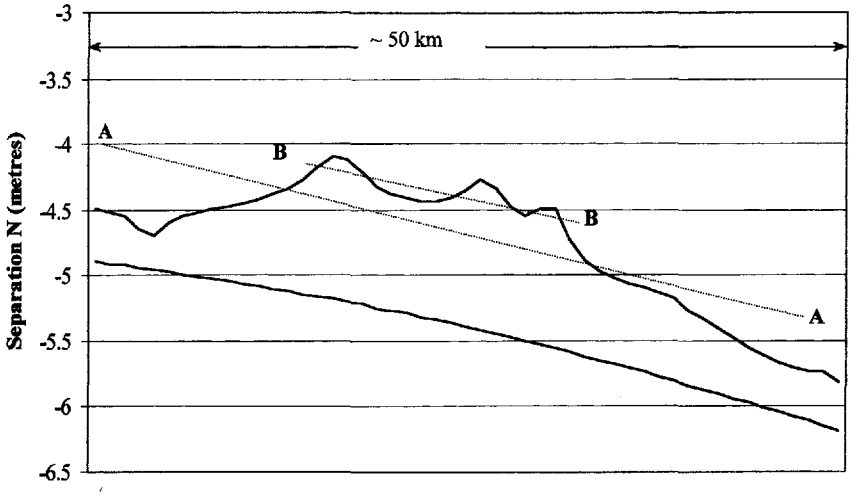

Figure 3.4 50 km cross-section of the geoid at latitude 16° S.

To answer the question, let us look a little closer at a smaller part of the geoidal cross-section shown in Fig. 3.3. This time, Fig. 3.4 shows a section around 50 km long. Looking at section AA in this example it seems that the geoid could be modelled by a uniform slope with a precision of around 50 cm for a distance of 50 km. For the shorter section BB, which is around 20 km, the precision is around 10–15 cm. We could tentatively infer that these results are typical, although it should be noted that the terrain in this test area is fairly rugged (with heights up to 1000 m), and a better result might be expected in less hilly terrain.

3.2 Reference surfaces for height

An important consequence of the irregularity of the geoid is that it leads to an alternative definition of height to the spheroidal height defined in section 2.3. This is the height of a point above the geoid or the *orthometric height*, usually given the symbol H.

As the height above the geoid is approximately the same as the height above mean sea level, this is the value that is usually used in surveying and mapping, and which is displayed on almost all maps and charts. It is, in fact, the height which can be obtained by levelling, provided that for very precise applications suitable corrections are made to counter the effect of convergence of the Earth's equipotential surfaces – these corrections are measured in centimetres over many tens of kilometres.

The reference point for mean sea level in a national mapping system is usually a tide gauge. Since there is a difference between the geoid and mean sea level, as mentioned in section 3.1, the datum for height is slightly different in each country. The significance of this point is illustrated by considering a project such as the construction of the Channel Tunnel. If the tunnellers from each end were due to meet at a midpoint defined as a certain height above (or in this case below) the geoid, then in principal there would not be a problem since the geoid is a unique surface. In practice, however, the vertical reference surface in each country would be defined with respect to different tide gauges. To complicate matters further, these tide gauges are not placed on either side of the Channel (in which case a fairly small slope of the sea surface might be expected over this short distance) but in Newlyn (in Cornwall) and Marseilles (on the Mediterranean coast). The difference in the vertical datums between Great Britain and France is therefore a function of the slope of the sea surface over the rather longer (sea) distance between these two ports, and in fact amounts to some 30 ±8 cm (Willis *et al.*, 1989).

To emphasise what is happening in this situation, let us restate the definition of the geoid given in the previous section: *it is that equipotential surface that most closely corresponds to mean sea level.* In effect, we can now see that this is an ideal, and the practical realisation of the situation is in fact that each country has selected a slightly different equipotential surface. Rapp (1994) quantified these differences by comparing determinations of the geoid separation that were found from the differences between orthometric and ellipsoidal heights on the one hand, and from global geopotential models on the other. Although the precision of these determinations is not sufficient for high accuracy engineering work, they are illustrative of the general problem, and are summarised in Fig. 3.5.

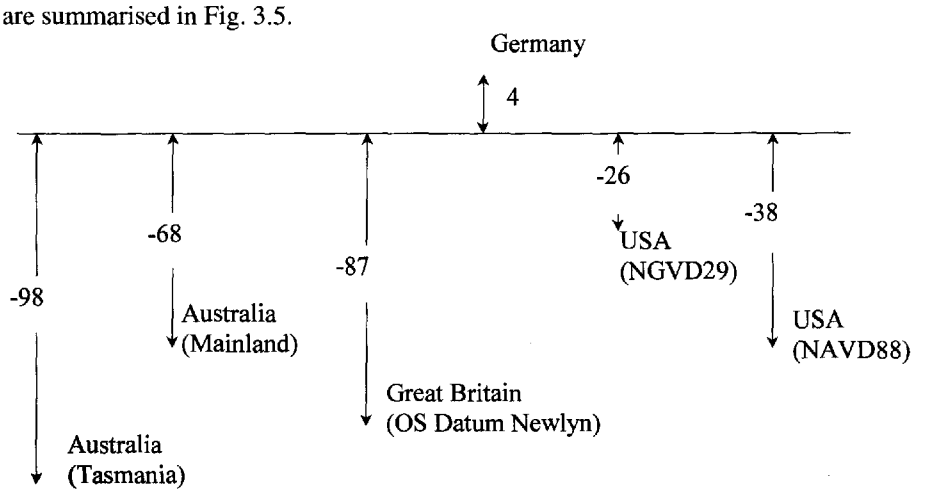

Figure 3.5 Reference surfaces with respect to the ideal geoid (after Rapp, 1994). All units are centimetres.

In some countries, the situation depicted above is itself an oversimplification. One example is Australia, where the vertical datum AHD71 (Australian Height Datum 1971) was established by adjusting a network of levelling observations across the country while holding the values fixed at 30 tide gauges spread around the coast (Hig-

gins *et al.*, 1998). Since each of these tide gauges is establishing the datum at a different equipotential surface, what is effectively happening is that the reference surface is no longer a single equipotential surface, but changes gradually from place to place. This is illustrated symbolically in Fig. 3.6. Over the kind of distances involved, this has not previously been a problem, as the differences were exceeded by the accuracy that could be achieved from levelling. It does present a problem, however, in quoting a value for the geoid–spheroid separation that should be used to correct heights determined from GPS. This is another example of the way in which datums are having to be redefined in the light of the increased accuracies that are possible with satellite systems.

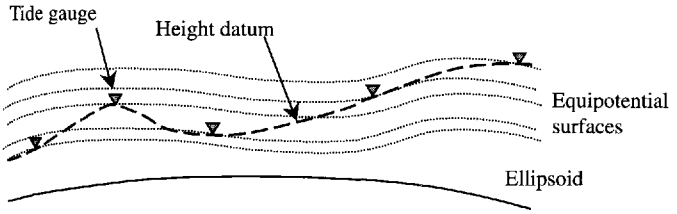

Figure 3.6 Vertical datum tied to several tide gauges.

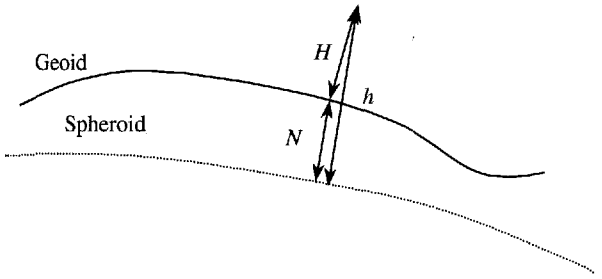

Figure 3.7 Orthometric height and spheroidal height.

The relationship between the orthometric height, H, and the spheroidal height, h, is as shown in Fig. 3.7. Since the angle between the vertical and the normal is so small, it can be said without any significant loss of accuracy that:

$$h = H + N \tag{3.1}$$

It is emphasised that the height h is the value that normally results from satellite observations; the significance of this equation is therefore that in order to obtain a height above the geoid, H, it is necessary to know the separation, N.

Alternatively, since most precise satellite positioning is done by relative techniques (as described in Chapter 5), (3.1) can be recast in the form:

$$\Delta h = \Delta H + \Delta N \tag{3.2}$$

This means that a satellite system such as GPS can be used to find the change in spheroidal height (Δh) between, say, a bench mark and a new point. In order to find

the orthometric height of the new point above the bench mark (ΔH) it is necessary to know only the *change* in the geoid separation (ΔN). Within the accuracy limitations discussed in section 3.1, this can sometimes be ignored.

As a final comment here on reference surfaces for height, it is worth pointing out that the advantages of using ellipsoidal heights should not be overlooked. Although these are not always appropriate, there are many situations in which the physical link with the direction of gravity, and with a surface perpendicular to this, is unnecessary. An example of this might be the monitoring of the movement of a point over time, perhaps one near the summit of a volcano to give an indication of future eruption. In this situation, a *change* in the height is all that is required, and a change in an ellipsoidal height is the same as the change in the orthometric height to any accuracy that could conceivably be required. The need for a determination of the separation is thus avoided, as is the danger of an apparent change being recorded as a result of an improved determination of the separation at a later date.

To take another example, an aircraft landing at an airfield and using a satellite positioning system (giving ellipsoidal height) can determine its height above obstructions that have themselves been surveyed using satellite techniques and have their heights expressed in ellipsoidal terms. In this context, the position of the mean sea level surface becomes an abstract and irrelevant concept.

What might be thought of as a halfway position between using orthometric heights and ellipsoidal heights is to adopt a standard model of the geoid to correct satellite observations: if the model is not perfect, true orthometric heights will not be obtained. On the other hand, if the use of the model is universal only the most precise engineering applications will need to use any other form of height coordinate, and the advantage will be that a consistent system will be adopted.

4
Global, regional, and local datums

4.1 Global datums

4.1.1 Satellite datums

On a global basis, the most appropriate version of a spheroid is one that has its origin at the centre of mass of the Earth and is of such a shape and size that it is the best possible approximation to the form of the geoid. Such a spheroid is necessary for any worldwide application, and is therefore the basis of a satellite reference system or datum. A datum with its origin at the centre of the Earth is said to be *geocentric*. This is a necessary condition for computations of satellite orbits. Since the advent of satellite geodesy in the early 1960s, there have been several attempts to define a global geocentric datum. The most recent of these is the World Geodetic System 1984 (WGS84).

The parameters of WGS84 are defined (Seeber, 1993) as :

$$a = 6\,378\,137 \qquad f = 1/298.257\,223\,563$$

as well as further parameters relating to the gravitational field and the rate of rotation. Note that the shape and size of the spheroid thus defined are almost consistent with another commonly used global datum, the Geodetic Reference System 1980 (GRS80), which is mainly used in gravitational applications. In fact, due to a slight difference in the original terms used for the definition of these two datums, there is a difference in their flattening (Hofmann-Wellenhof *et al.*, 1997) that amounts to :

$$\Delta f = f_{\mathrm{GRS}} - f_{\mathrm{WGS84}} = 16 \times 10^{-12} \tag{4.1}$$

For most practical purposes, this difference can be ignored and the spheroids considered equivalent.

The concept of realising a reference system, as opposed to defining it, is one that has already been touched on with respect to the vertical height datums of the last section, and will be explored more fully in the section on local and regional datums that

follows. It is also something that should be considered when dealing with a satellite datum. In effect, this means that, although WGS84 can be defined as a geocentric datum and related to a certain spheroid, most users will encounter this reference system through the use of the GPS (Chapter 5). The starting point for their positioning is the satellite ephemerides (or orbital positions) that are broadcast by the system, which in turn have been determined from a set of monitoring stations. It is these monitoring stations that effectively realise the coordinate system. Using a different set of monitoring stations to determine the ephemerides (as, for example, the far more extensive network of the International GPS for Geodynamics Service) would then lead to a separation of the datum into two separate ones. This problem has now been largely overcome by linking all important global networks into the International Terrestrial Reference Framework discussed in the next section.

An earlier attempt to define a geocentric reference system was WGS72. The parameters of this are:

$$a = 6378135 \qquad f = 1/298.26$$

WGS72 was a less accurate system than WGS84, and its origin has now been shown to miss the centre of the Earth by almost 5 m. Figure 4.1 shows the main differences between the two systems.

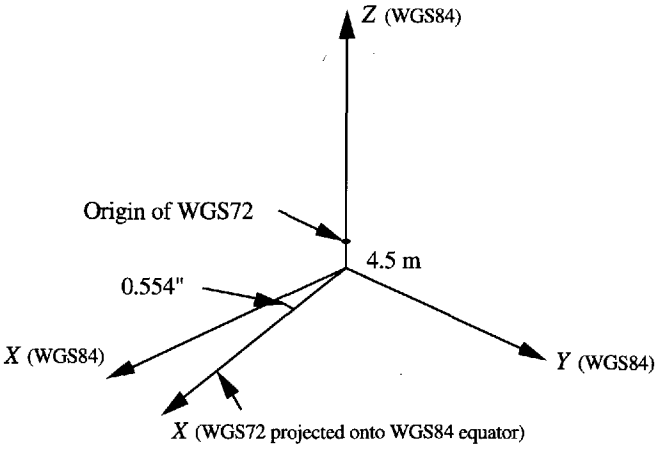

Figure 4.1 Principal differences between WGS84 and WGS72.

WGS72 is no longer used for the acquisition of new data, so its interest is really only in the interpretation of archive material. Again, it should be noted that there were several different realisations of WGS72 resulting from the use of broadcast and precise ephemerides.

Another group of satellite datums are those associated with GLONASS (Global Navigation Satellite System), the Russian equivalent of GPS. The datums used for this are the Soviet Geodetic System 1985 (SGS85) and its successor SGS90 (introduced in 1995, and alternatively referred to as PZ90). The defining geometrical parameters of both SGS85 and SGS90 (Hooijberg, 1997) are:

$$a = 6378136.0 \qquad f = 1/298.257$$

The shape and size of SGS90 are therefore very similar to those of WGS84. Several attempts have been made to determine the transformation parameters between the two, such as those by Misra and Abbot (1994) and Rossbach *et al.* (1996). These have generally given similar results within the accuracy limitations that have been imposed by the use of relatively small networks of points used.

The best current indications are that PZ90 has a zero meridian that is rotated 0.6″ to the east with respect to that of WGS84, and an equatorial plane that lies around 4 m to the north of WGS84's. These discrepancies would show up directly between absolute determinations of position by the two systems. In terms of relative positioning between two receivers, the rotation of 0.6″ is the only relevant parameter, and is equivalent to 1 cm over a distance of about 3.5 km. If a receiver makes use of satellites from both systems, the differences in the datums is not a problem for short distance or low accuracy applications; for high-precision baselines over long distances, better transformations will have to be developed (Walsh and Daly, 1998).

4.1.2 International reference frames

An international terrestrial reference system (ITRS) is one that is made use of throughout the world: WGS84 and PZ90 are examples of these. The most fundamental coordinate datum in existence is the International Terrestrial Reference Framework (ITRF). It is a highly accurate geocentric reference framework, with the coordinates of reference stations defined at the millimetre level.

The preparation and evaluation of the ITRF is the responsibility of the International Earth Rotation Service (IERS), which was established in 1998 to replace the International Polar Motion Service and the Earth rotation functions of the Bureau International de l'Heure. The wider remit of the IERS is to establish both the terrestrial reference frame and the international celestial reference system (a coordinate system that does not rotate with the Earth), and to determine the connections between the two systems. This process involves monitoring and predicting the rotation of the Earth and the movement of its poles, and is essential information if, for example, a determination of a satellite's position is to be given in a terrestrial system such as WGS84.

For the level of accuracy involved, it is no longer possible to have a reference system that is independent of time, since the deformations of the Earth's crust mean that the coordinates of points within the system are continually moving, typically by several centimetres per year. In some respects, extremely high precision positioning is similar to very slow moving navigation.

The ITRF includes the velocities of its stations in the definition of the datum, but applications making use of its coordinates will usually do so by reference to a particular epoch, such as ITRF92, ITRF93, and ITRF94. The latest solution, ITRF96, has determined the positions at 290 stations worldwide, of which 40% have position uncertainties below 1 cm (IERS, 1998).

WGS84 was previously only defined at an accuracy level of around 50 cm. It was redefined in 1994, however, in a form that was compatible with ITRF92, and strictly speaking was then referred to as WGS84 (G730), with 730 being the GPS week number in which the change was effected (the first full week of January 1994). Hence WGS84 in this form is compatible with ITRF92 at the level of 10 cm (Hooijberg,

1997). There have been other redefinitions since then, such as WGS84 (G873), and although these are all slightly different datums this is noticeable only for applications of the very highest accuracy.

The 290 stations of the ITRF are rather thinly spread around the world, and for practical purposes it is necessary to densify this network: it can also be inconvenient in some situations to have a coordinate set that is continuously changing. The European Terrestrial Reference Framework 1989 (ETRF89) is a subset of ITRF defined in the European region, and was equivalent to the ITRF at the time of its implementation. There is in consequence a mismatch between ETRF89 and the current epoch of the ITRF. This becomes apparent, for example, if points are being fixed with respect to permanent GPS positions whose coordinates are published in ITRF, but where the final object is to position points in the ETRF system. The discrepancy amounts to several centimetres per year since ETRF was established, and a transformation is required for high precision applications. The transformation parameters can be obtained from, for example, Boucher and Altamimi (1998). Comments on the link between this system and the British one are given in section 4.2.2.

4.2 Local and regional datums

4.2.1 Definition

A datum is defined by selecting an origin for a national or regional survey. At this point the geoid–spheroid separation, N, and the deviation of the vertical are chosen, usually as zero. This has the effect of fixing the chosen spheroid to the geoid (coincident and parallel) at the point of origin. The orientation of the minor axis of the spheroid is made parallel to the spin axis of the Earth by making *Laplace azimuth observations*: observations of astronomical azimuth (with respect to the stars and the pole of the Earth's rotation) are converted to geodetic azimuth (defined on the ellipsoid) by a formula that forces the poles of the two systems to be the same. This combination of *shape* and *size* as given by the spheroid, and *position* as given by the fixing at the origin, is essentially what defines a datum.

A spheroid is *not* a datum. Many countries may use the same spheroid, but they are on different datums as they have different points of origin. Before the advent of satellite geodesy, all national datums were of necessity defined independently of each other.

The effect of defining the datum in this way is that the spheroid is not geocentric. Figure 4.2 shows the relationship between a local datum and WGS84. It can be seen that in general a point has different coordinates in the two reference systems, whether these are expressed in geodetic or cartesian form.

Almost by definition, a local datum approximates the geoid in the region much more closely than does the global datum, or a datum optimised for a wider region. This point is illustrated by Figs 4.3 and 4.4, which show the geoid with respect to the British and European datums respectively. Note that in these two diagrams the geoid is the same; it is the reference system which has changed. Each local or regional datum therefore has a point of origin which is offset from the centre of the Earth. The size of this offset may be as much as 1 km. It is usually expressed in components

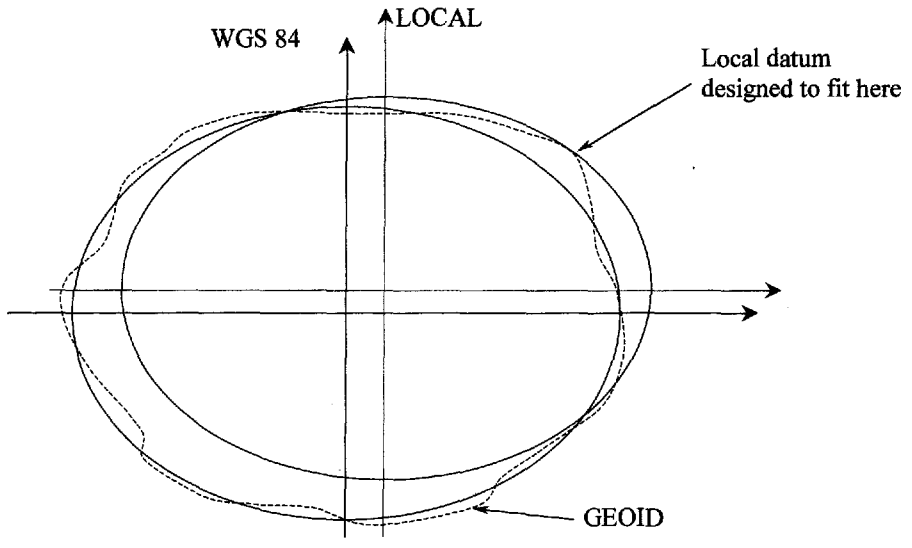

Figure 4.2 A local datum and WGS84.

$(\Delta X, \Delta Y, \Delta Z)$, where these values are effectively the coordinates of the origin of the local datum in WGS84. Hence a simple transformation of coordinates from a local system to WGS84 is given by:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{\text{WGS84}} = \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} + \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{\text{local}} \tag{4.2}$$

The transformation from geodetic coordinates on one datum to geodetic coordinates on another can thus be achieved by the intermediate step of first converting to cartesian coordinates via equations (2.5), applying equation (4.2), and then converting back to geodetic coordinates via equations (2.7)–(2.12).

Alternatively, it is possible to combine all these steps in one set of equations, such as the Molodensky formulae (Stansell, 1978). These give the changes in geodetic coordinates directly as:

$$\Delta\phi = \frac{\left\{ \begin{array}{l} -\Delta X \sin\phi\cos\lambda - \Delta Y \sin\phi \sin\lambda + \Delta Z \cos\phi \\ \quad + \Delta a[(\nu e^2 \sin\phi\cos\phi)/a] + \Delta f[\rho(a/b) + \nu(b/a)]\sin\phi\cos\phi \end{array} \right\}}{(\rho + h)\sin\lambda} \tag{4.3}$$

$$\Delta\lambda = \frac{-\Delta X \sin\lambda + \Delta Y \cos\lambda}{(\nu + h)\cos\phi \sin\lambda} \tag{4.4}$$

$$\Delta h = \Delta X \cos\phi\cos\lambda + \Delta Y \cos\phi \sin\lambda + \Delta Z \sin\phi - \Delta a \frac{a}{\nu} + \Delta f \frac{b}{a} \nu \sin^2\phi \tag{4.5}$$

where all terms have previously been defined except

$$\rho = \frac{a(1-e^2)}{(1-e^2\sin^2\phi)^{3/2}} \tag{4.6}$$

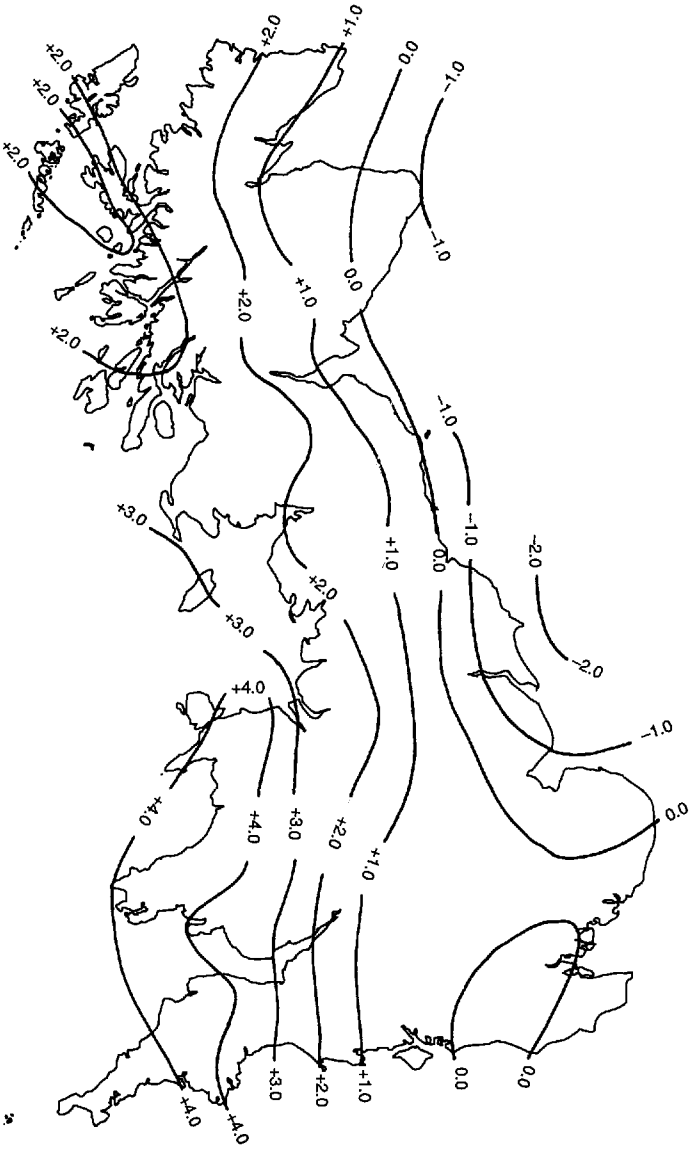

Figure 4.3 The geoid with respect to the British datum, OSGB36. (Courtesy of J. Olliver, Oxford University.)

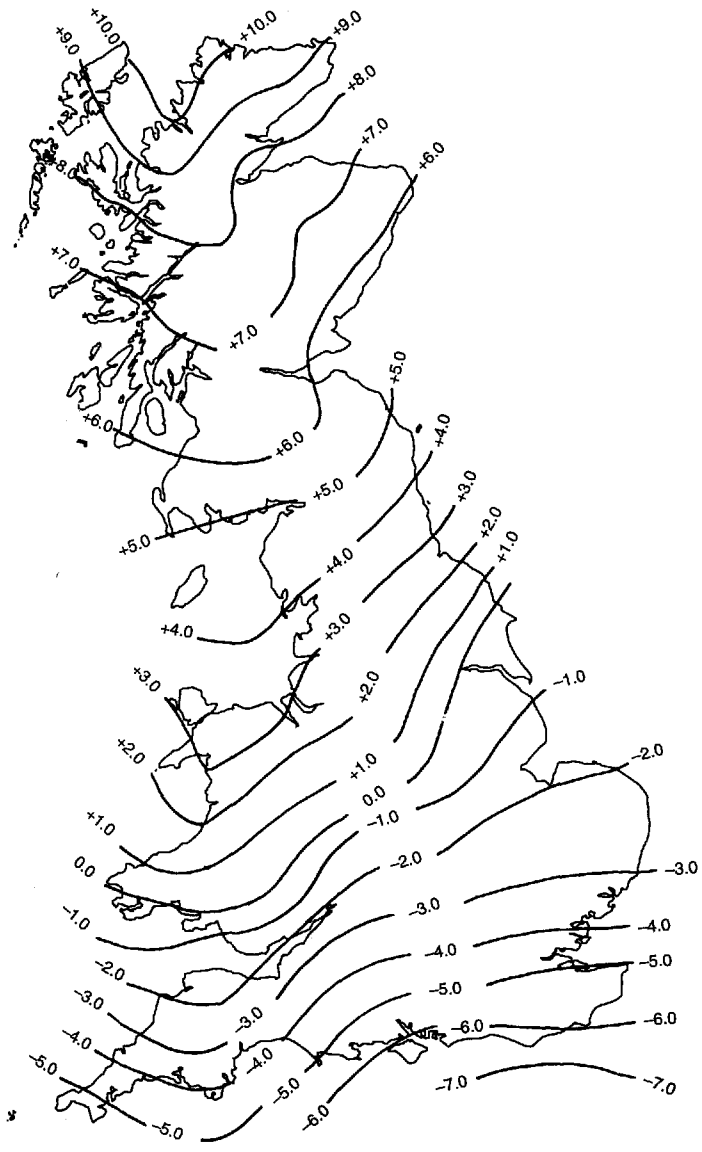

Figure 4.4 The geoid with respect to the European datum, ED50. (Courtesy of J. Olliver, Oxford University.)

4.2.2 Realisation of a datum

The previous section provided a definition of a reference system. In fact, before it can be used it is necessary to *realise* the datum; that is, to provide physical monuments of known coordinates. Any mapping or surveying which is carried out is then related to the coordinates of those control points that are available locally, rather than to the rather abstract concept of a datum defined at an origin point.

This then presents the problem that the datum as realised includes all the measurement errors and computational approximations of the original survey, which would usually have been carried out by a process of triangulation using ground survey techniques. Another significant effect is that the minor axis of the spheroid would have been made parallel to the spin axis of the Earth as it existed at the time of the original survey; in fact, the 'wander' of the pole means that modern datums use a conventionally agreed point, rather than the pole's instantaneous position at any one epoch. It may also be the case that the transformation is between two datums that are *both* distorted in this way: Appendix section A3 covers the mathematical treatment of this.

It is therefore necessary to modify the rather simple transformation procedure outlined in section 4.2.1 in two ways. Firstly, to accept that in addition to the simple translation between two datums there may also exist rotations and scale changes. This leads to a modified set of transformation equations:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{\text{WGS84}} = \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} + \mu \mathbf{R} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{\text{local}} \tag{4.7}$$

where μ is the scale factor between the two systems and $\mathbf{R}$ is a rotation matrix, which for small angles α_1, α_2, α_3 about the X, Y, Z axes (as shown in Fig. 4.5) can be expressed as:

$$\mathbf{R} = \begin{pmatrix} 1 & \alpha_3 & -\alpha_2 \\ -\alpha_3 & 1 & \alpha_1 \\ \alpha_2 & -\alpha_1 & 1 \end{pmatrix} \tag{4.8}$$

The transformation described by equation (4.7) is not the only way of carrying out a three-dimensional transformation, and in many ways it is not the best. An alternative is given in section 6.5, and a more comprehensive mathematical treatment is given in Appendix section A3.

The second modification relates to the distortions in the local datum. The amount of distortion involved will be dependent on the size of the country (larger surveys accumulate more errors away from the origin) and the way in which the original survey and computations were carried out. A typical figure for the distortion of a country such as Great Britain would be around 20 m (that is, the difference between the actual distance between two points at either end of the country and the distance implied by their coordinates).

It is possible to regard the resulting situation in two different ways. Firstly, the transformation given by equation (4.7) will convert coordinates into the datum, but these will then be different from the ones that are actually used, owing to the 'errors' in the local survey. It is important, however, to then add these errors to the converted

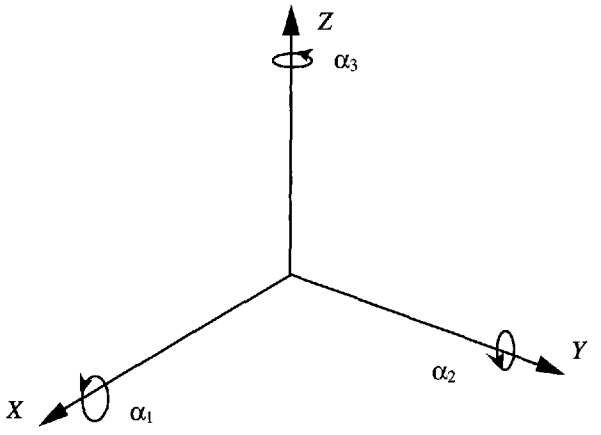

Figure 4.5 Rotation convention.

coordinates to bring them into line as *the errors are part of the datum.* That is, once the error has been made it is important for consistency that all users adopt it. This is not quite the chaotic situation it implies, because in general the errors at points in any one region of the survey will be very similar to each other (although in Great Britain some quite sharp discontinuities do exist because of the way in which the original survey was computed). This is effectively the approach used by the Ordnance Survey, which has published a method for determining the additional distortions to an accuracy of around 2 m (Ordnance Survey, 1995a). The Ordnance Survey also provides a service that will determine the distortions to an accuracy of 20 cm. This is discussed further in section 6.5.

An alternative way of looking at the problem, and one that is generally of more use to those determining their own transformation parameters, is to regard a small region of the local system as largely undistorted (since the errors in that region are highly correlated with each other), and then to derive a set of transformation parameters as in the model given in equation (4.7). These will be applicable only to the region for which they were derived, and will not in general be applicable to other areas. Thus, conceptually, a situation arises where the transformation parameters can be thought of as variables that change across the country. It is even possible to produce a set of maps that show, for example, how the ΔX parameter varies, although this is liable to lull the user into a false sense of the exactitude of such a procedure.

Whichever of these methods is used to conceptualise or to solve the problem, it must be emphasised that no one set of transformation parameters on its own is likely to be sufficient except at accuracies of a few tens of metres. Figures quoted for the datum of any particular country will be average values and not applicable throughout the country without some loss of accuracy.

Some examples of transformation parameters for various datums are given in Table 4.1; the values given here are averages, and should be used with caution. Notes on sourcing this information are given in section 6.4.1.

The datum transformation parameters for adjacent countries are correlated only if there was a political link at the time of the establishment of the original survey, as

Table 4.1 Examples of transformation parameters from local datums to WGS84

Datum name	Spheroid	ΔX (m)	ΔY (m)	ΔZ (m)
OSGB 36	Airy	−375	111	−431
Ireland 1965	Airy 1830 Mod	−506	122	−611
French	Clarke 1880 (IGN)	−168	−72	314
German (West)	Bessel 1841	583	68	395
Netherlands	Bessel 1841	593	19	468
Swiss	Bessel 1841	679	0	406
European 1950	International 1924	87	98	121
Cape (South Africa)	Clarke 1880 Mod	136	108	292
Arc 1950 (Zimbabwe)	Clarke 1880 Mod	142	96	293
Namibian	Bessel 1841 Nam	−616	−97	251
North American 1927 (Conterminous USA)	Clarke 1866	8	−160	−176
North American 1927 (Canada)	Clarke 1866	10	−158	−187
Indian (India, Bangladesh, Nepal)	Everest 1830C	−289	−734	−257
Timbalai 1968 (Sabah)	Everest 1830B	689	−691	46
Tokyo (Japan)	Bessel 1841	123	−483	−662

Table 4.1 shows. Thus, the parameters for South Africa and Zimbabwe are similar, whereas those for Namibia (formerly German South West Africa) are very different.

There is a trend in some countries (for example the USA and Australia) for the ITRF to be adopted as the basis for the national mapping system. The advantages ensuing from such a situation must be balanced against the costs of converting all existing data to the new datum.

In Great Britain (but not in Northern Ireland) the official datum is OSGB36; the coordinates of all triangulation points and all maps are based on this system. Alongside this there is a datum known as OS(GPS)93, which is specifically for use with GPS (Chapter 5). This datum is entirely compatible with ETRF89, and there is hence a rather complex transformation between this and OSGB36. The possibility of adopting ETRF89 as the basis for mapping in Great Britain at some future date is being kept under review (Calvert, 1994), but at present the emphasis is on promoting the acceptance of a consistent set of conversion algorithms (Ordnance Survey, 1997, 1998).

5
The global positioning system

5.1 Introduction

For many users of spatial data, the idea that maps can be on different datums, and indeed on a different one from the method of data acquisition, will first have been encountered when using the global positioning system (GPS). For this reason, although it would be possible (and quite reasonable) to write an entire book devoted to this system, it is appropriate to summarise its most important features here, and to explore some of the datum aspects in more detail.

A more extensive treatment can be found in standard texts such as Hofmann-Wellenhof *et al.* (1997) or Leick (1990).

5.2 System overview

The GPS was originally conceived as a navigation system – that is, one capable of instantaneous determinations of position and velocity. As will be seen, the minimum requirement for its operation is the ability of a user to receive signals from at least four satellites simultaneously: for locations that are uncluttered by obstructions in the form of buildings or dense foliage this requirement is assured by the constellation of 21 satellites at a mean height of 20 200 km above the surface of the earth, as shown in Fig. 5.1.

The basic premise of GPS is the same as that of any surveying system: the coordinates of new points are found by making observations with respect to points of known coordinates. The only differences here are that the known points are in orbit, and are not stationary. The determination of the coordinates of these satellites is therefore a continuous process, and is achieved by the *control segment* of GPS, which consists of a worldwide network of monitoring and control stations dedicated to the task of determining the orbital paths of the satellites and monitoring the health of their signals. It is then possible to predict the orbit of a satellite a short way into the future, and to upload this information to the satellite itself. In this way, the satellite is able to broadcast its position (referred to as the *ephemeris*) for the users to determine their own position in real time.

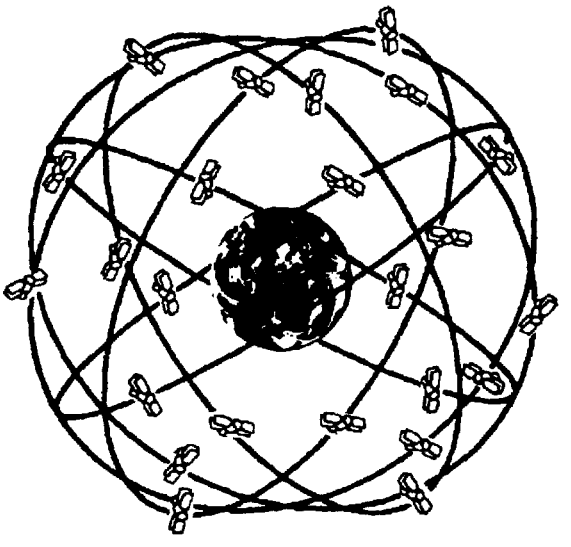

Figure 5.1 The GPS constellation.

The information transmitted by the GPS satellites is modulated onto two underlying carrier waves, referred to as L1 and L2. The former has a frequency of 1575.42 MHz (equivalent to a wavelength of 19.05 cm) and the latter has a frequency of 1227.60 MHz (equivalent to 24.45 cm).

In addition to the ephemeris and other general information about the system, two binary codes are also modulated onto the carrier waves. These can be thought of as being the equivalent of the step function shown in Fig. 5.2, with each change in sign being represented by a change in phase of the carrier wave. The two codes are referred to as the *coarse acquisition (C/A) code* and the *precise (P) code*, and their effective wavelengths are 300 m and 30 m respectively. Although the P code is publicly available, a further encryption (the W code) is added which makes it inaccessible to most users. The resulting combination is referred to as the Y code, and the procedure of implementing this encryption is known as *anti-spoofing*. In fact, the C/A code is modulated onto the L1 carrier only, whereas the P code is modulated onto both frequencies, a configuration that deliberately denies the non-military user the advantages of a dual frequency system (although this is overcome for some applications).

Figure 5.2 Binary code carried on the GPS signal.

The main point to note about the codes at this stage is that they are transmitted in a predetermined sequence at precise times for each satellite. Thus, if the codes can be read, they can simply be thought of as time information.

5.3 Positioning with codes

The basic method for positioning with GPS is through the use of the codes, usually only the C/A code on the L1 carrier wave. The procedure is for the receiver to 'read' the code, and thus obtain the time information that is transmitted by the satellite. This procedure is aided by the fact that the receiver has a copy of the code stored within it, and it knows roughly what to expect: it is therefore possible to make a reading on what is in fact a very weak signal (compare a handheld GPS receiver with a satellite television aerial).

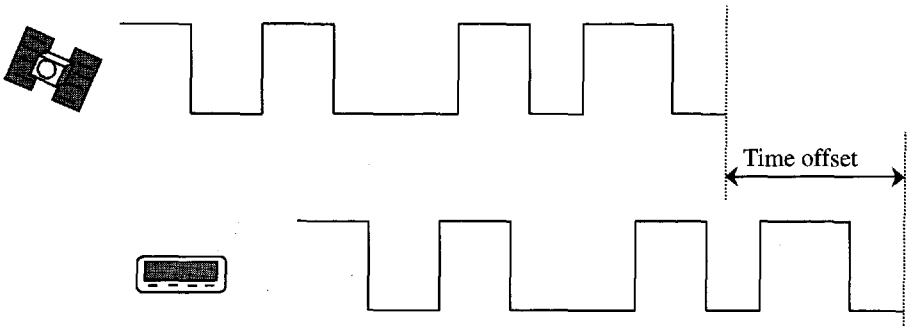

Figure 5.3 Time shift between codes.

The receiver, generating the code itself, determines the offset in time between the code that it generates and the one received from the satellite, and from this deduces the time of travel of the signal. By multiplying this travel time by the speed of transmission of the signal (the speed of light), the distance from the satellite to the receiver is obtained. In theory, it would be possible to determine the three-dimensional coordinates of the receiver from observations to three satellites. This follows either from considering three simultaneous equations with the three unknowns (X,Y,Z), or from a geometrical analogy in which each determination of a distance describes a sphere of a certain radius centred on the satellite, and the intersection of three spheres is a single point.

The model described above has taken no account of the lack of synchronisation between the satellite time system and the receiver clock, however. Although the synchronisation between the different satellites is essentially assured by each one carrying two rubidium and two caesium atomic clocks, and these being continuously monitored by the ground stations, the receiver has to have an independent, and much cheaper, timing system. This then leads to the problem that, because the speed of light is approximately 300 000 km/s, a 1 s offset between the satellite and receiver clocks will lead to a 300 000 km error in the determination of the distance. For this reason, the distances so found are sometimes referred to as *pseudo-ranges*. The problem is solved simply by introducing a fourth satellite, and using four distances to solve the four unknown parameters (X,Y,Z, $\Delta\tau$) where $\Delta\tau$ is the offset between the GPS system time and the receiver's clock.

The reasons for the design of the satellite constellation are now apparent: four satellites is the minimum requirement, and in fact this is virtually always possible in

the absence of any obstructions (it should be noted that the signals can travel through cloud, fog, rain, and so on, but cannot penetrate buildings, dense foliage, or similar bodies). Usually between four and eight satellites will be above the horizon.

The determination of a satellite–receiver distance from code observations is subject to several sources of error. These can be summarised as follows, with approximate root mean square (rms) magnitudes as given by Langley (1997):

- *Ephemeris error:* caused by the difference between the satellite's true position and the broadcast ephemeris, and in which can also be included the errors in the satellite clock. This will usually amount to around 3–4 m.

- *Refraction:* caused by the difference between the actual speed of transmission and its assumed value. This error source can be subdivided into two components: ionospheric and tropospheric refraction. The former has its origin in the upper atmosphere, and is very difficult to predict. It can cause errors of several metres in the observed range, but can be corrected by using dual frequency observations: this option is not available to civilian users, and the errors can be around 7 m. Tropospheric refraction, on the other hand, is more predictable. It refers to the refraction encountered in the lowest 40 km of the atmosphere, and is dependent on temperature, pressure, and humidity. The refractive index of the dry atmosphere is very easily modelled, and its effects are negligible: only the wet part of the atmosphere causes any real problems, and is typically less than 1 m.

- *Multipath:* caused by the satellite signal reflecting off other surfaces on its way to the antenna, and thus taking an indirect route: it can be minimised by good site selection. Its magnitude is dependent on the environment at the receiver, and is typically 1–2 m.

- *Receiver noise:* a random error source that is a reflection of the measuring precision of the receiver. A figure of 1.5 m for the rms error is typical.

To all of these error sources, an additional category can be added: a deliberate error, known as *selective availability*, which limits the accuracy that can be achieved by users of the C/A code. This is in theory composed of *delta* errors (fluctuations introduced in the satellite clock) and *epsilon* errors (a deliberate inaccuracy of the broadcast ephemeris) although indications are that at present the latter is not implemented. In combination with the 'natural' errors, selective availability limits the accuracy to 100 m in horizontal position, and to 150 m in height. These figures refer to the limits that will not be exceeded by 95% of the observations.

The nature of the selective availability is that the errors change only very slowly with time, and it therefore takes at least an hour before any averaging effect will reduce the errors. To achieve an absolute accuracy of 5–10 m takes around 8 h of observations. Figure 5.4 shows a typical plot of the position determined at a stationary point over a period of 1 h.

One further point should be noted relating to the accuracy of GPS, and that is the way in which the geometry of the satellites affects the final accuracy of the determination of the coordinates of the receiver. Essentially, if the distances measured

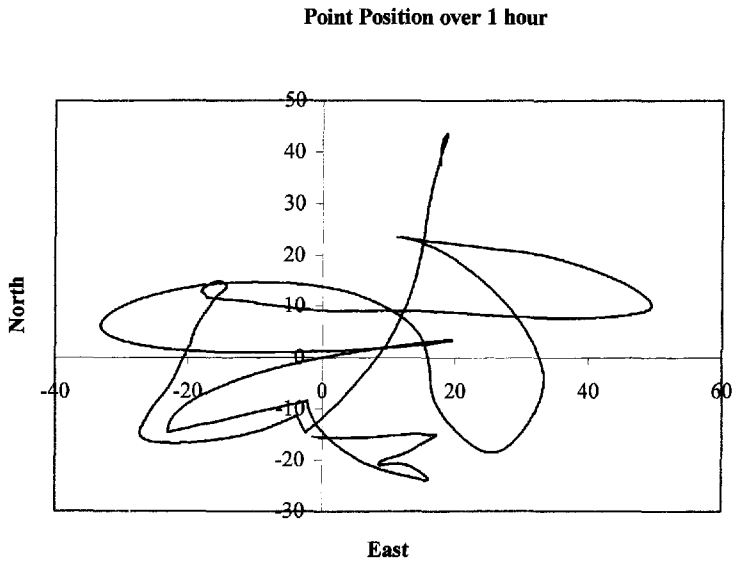

Figure 5.4 Point position over 1 h using code pseudo-ranges (SA on). (Courtesy of R. Keenan, University College London.)

from the satellites to the receiver intersect at a shallow angle (because the satellites are grouped in one portion of the sky), the accuracy of the three-dimensional coordinates of the point will be worse than if the satellites were well distributed around the sky. This is analogous to the situation with the determination of two-dimensional coordinates from the intersection of measured distances that is depicted in Fig. 5.5: although the distances are measured to the same precision in each case, the accuracy of the coordinates will be worse if the distances intersect at an acute angle. For GPS, this concept is conveniently expressed by a numerical coefficient, known as the positional dilution of precision (PDOP). This is the ratio of the mean accuracy of the coordinated position to the accuracy of the original range observations: the larger this number, the worse the geometry of the satellites. Typically, the PDOP values when using the full GPS constellation range between 2 and 5: if some satellites are obscured and the PDOP rises much above this range, the accuracy figures quoted above will no longer be achievable.

This, then, is the basis of operation of all GPS receivers operating as stand-alone units. It can be seen that the accuracy is almost entirely dependent on factors external to the receiver: any difference in price between different models is therefore explained by the functionality of the equipment, such as the storing of data, the use of digital map displays, and so on.

Future developments in the design of GPS satellites, and US government policy regarding the application of selective availability, mean that within a few years the accuracies achievable by stand-alone GPS receivers are likely to increase dramatically from the figures quoted here, possibly to as accurate as 5 m (GIAC, 1999).

If coordinates obtained by this method are to be integrated with data from a nation-

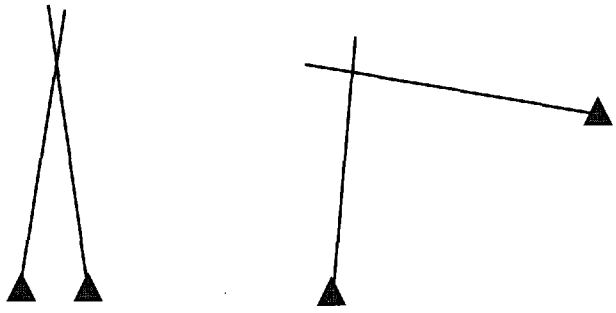

Figure 5.5 Poor geometry (left) and strong geometry (right).

al survey, a datum transformation must be carried out. The coordinates will change by up to 1 km after transformation, which is significant even for the current accuracy of GPS fixes. Distortions in the local datum are unlikely to be significant, however, and an average set of transformation parameters for the datum concerned can be applied. Most handheld receivers now on the market have stored within them a set of transformation parameters for the datums of many countries around the world. If these are not available, a simple shift can be derived by setting a receiver over a point with coordinates known in the local system, and then re-arranging equation (4.2) as:

$$\begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{\text{WGS84}} - \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{\text{local}} \tag{5.1}$$

The shifts derived in equation (5.1) can then be applied to all the other points that have been observed with GPS.

The height information that results from this transformation is the ellipsoidal height above the local datum. We saw in the example for the British datum (OSGB36) in Fig. 4.3 that this approximates the height above the geoid within 4 m. This accuracy is more than sufficient for this mode of operation.

5.4 Differential GPS using codes

The key to the improvement of the accuracy obtainable with GPS is the use of two or more receivers to measure *relative* as opposed to *absolute* positions. That this achieves an improvement in accuracy of almost two orders of magnitude is due to the spatial correlation of most of the errors: that is, an error present in the observation of range made at one receiver is very likely to be present at another receiver in the vicinity. This is certainly true of the errors caused by selective availability; the correlation of errors caused by the ephemeris and refraction will depend upon the distance between the two stations. Only multipath and noise are entirely uncorrelated between the two receivers.

The basic principle of operation is therefore to have one receiver set up at a point of known coordinates. In this context, 'known' implies known in the WGS84 system,

as local coordinate systems are entirely inappropriate for this type of computation. The information could be obtained by the mean of a long-term set of absolute position observations by the receiver or, more appropriately, by linking the point to a higher order reference system through the type of observation discussed in the next section. The effect of an error in the absolute coordinates of this station will be similar to the effect of an error in the satellite ephemeris when calculating a vector from one point to another; therefore, an appropriate target for the accuracy is better than 1 m.

The receiver at the known station, the *reference receiver*, is thus able to compare the observed pseudo-range to each satellite with the value calculated from its own coordinates and the ephemeris position of the satellite. From the difference between the two, a range correction can be determined (Fig. 5.6). This range correction can then be used to correct the range observations made at another receiver, sometimes referred to as the *rover*. The efficacy of this correction will depend on the degree of correlation between the errors at the two receivers. For example, an error in the satellite ephemeris will not cause an error in the range measured at the reference receiver if the direction of the error is perpendicular to the line from the satellite to the receiver. In turn, this undetected error will not cause a range error at the roving receiver if it is also perpendicular to the range, but this will be less true as the inter-station distance increases. Similarly, the degree of correlation in the effects of atmospheric refraction will decrease as the receivers become further apart.

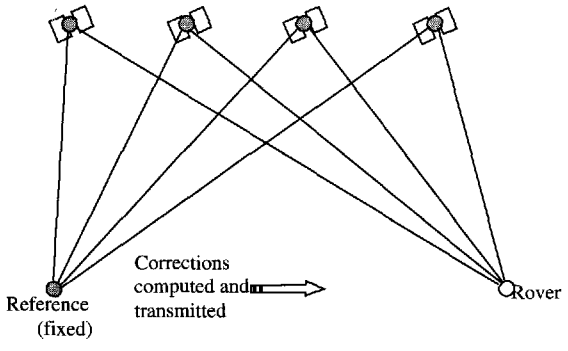

Figure 5.6 Differential GPS.

Generally, differential GPS (DGPS) will achieve an accuracy of 1–5 m, over distances up to several tens, or even hundreds, of kilometres. Figure 5.7 shows an example plot of the position of a static receiver over a period of 1 h, as computed over a baseline from a reference receiver of less than 1 km.

For many applications, the corrections to the observations made at the roving station can be made after the event, once the data from the two receivers has been downloaded onto a computer. This mode of operation is entirely suited to the determination of the coordinates of a series of static features for incorporation into a database. The alternative is to carry out the corrections in real time, in which case it is necessary to transmit the necessary information from the reference receiver to the rover. This can be done by purchasing two receivers and a radio communication link, in which case the reference receiver can be positioned wherever is appropriate for a particular

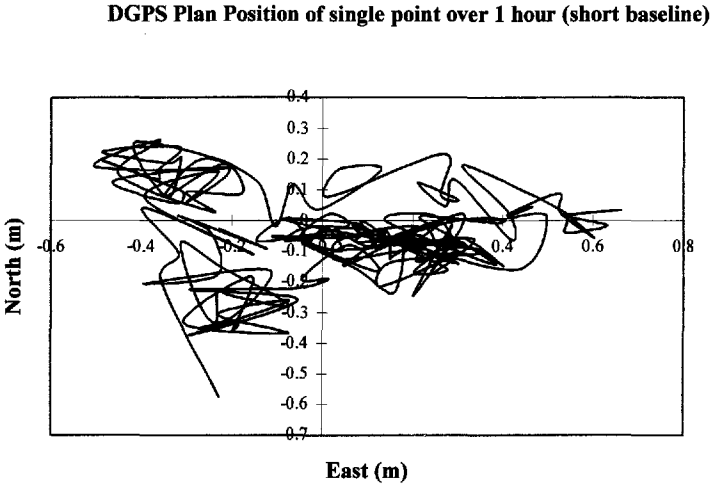

Figure 5.7 Coordinates determined by DGPS over a 1 h period. (Courtesy of R. Keenan, University College London.)

project. On the other hand, the distances over which DGPS is viable means that it makes commercial sense to establish a network of permanent reference stations and transmit the corrections to whoever is prepared to pay for them. The cost of the subscription to such a service is offset by not having to establish a reference station and a transmitter.

On land, the corrections can, for example, be transmitted on the side band of a commercial radio station, in which case the outlay of the user is limited to the purchase of an FM receiver in addition to the subscription. Alternatively, offshore users can receive corrections via a satellite communications link.

If the coordinates obtained from this type of survey are required in the local datum, rather than WGS84, they could simply be transformed to the local system *en bloc* using published values of the transformation from WGS84 to the local datum. This has the disadvantages of using general national values of the transformation, which are likely to be less accurate that the results obtained by DGPS. In addition, any errors in the WGS84 coordinates of the reference station will contribute directly as an additional error source. The alternative is to derive transformation parameters by occupying one or more points whose coordinates in the local system are known, and using equation (5.1).

If the survey is particularly extensive, however, it should be noted that rotations between the coordinate systems may be significant. A typical rotation might be 5 arc seconds, which is equivalent to 2.5 m over 100 km. This is unlikely to be a problem very often.

The height information yielded by this method is differential spheroidal height. The difference between this quantity and a differential geoidal height is barely sig-

nificant over short distances at the accuracy achievable. Over longer distances (say >50–100 km), adequate geoid information would be provided by a global earth model such as EGM96, which can correct the differential spheroidal height through the use of equation (3.2).

5.5 GPS phase measurements

The factor that limits the precision of GPS code observations is the effective wavelength of the code: at 300 m for the C/A code very precise observations are impossible. The way around this is to ignore the codes (they can be removed from the signal if they are known) and go back to the underlying carrier waves, L1 and L2, which have wavelengths of 19.05 cm and 24.45 cm respectively. This allows the possibility of very precise observations, as measuring a fraction of a wavelength offers millimetric precision – although at the cost of an increased mathematical complexity of the solution. Whereas the code observations represent unambiguous determinations of distance, the phase observations measure only a fractional part of the distance, which repeats itself every wavelength.

The basic observation is made by measuring the difference in phase between the signal received from the satellite and one generated in the receiver. The difference in phase so measured has two causes.

- The receiver and the satellite are not likely to be oscillating in phase in the first place.
- The signal has had to travel a certain distance from the satellite to the receiver, so the receiver is comparing its own signal with one emitted from the satellite a short time previously. If the satellite were a whole number of wavelengths away from the receiver, the two signals would once again be in phase with each other. Hence (Fig. 5.8), the receiver is actually measuring the fractional part of the distance $\Delta\phi$, over and above a whole number of wavelengths. Subsequent measurements in the same sequence can measure a value of $\Delta\phi$ as it goes beyond a whole wavelength, so that the initial unknown number of whole wavelengths stays the same.

The result is therefore an observation of phase which is related to the satellite–receiver range, but in a way that is complicated by the presence of an unknown whole number of wavelengths (the *integer ambiguity*) and phase differences between the satellite and receiver clocks. The key to the solution of this problem is the fact that these *bias terms* are constant, provided that the receiver continuously tracks the signal from the satellite. It is therefore possible, in time, to acquire more and more observations without increasing the number of unknown parameters to be solved.

There are several different approaches to solving the problem.

- A single receiver can occupy a point for a considerable period of time (at least several hours), collecting a very large data set. In combination with a more precise ephemeris than that broadcast by the satellites (for example the precise ephemerides disseminated via the International GPS for Geodynamics Service,

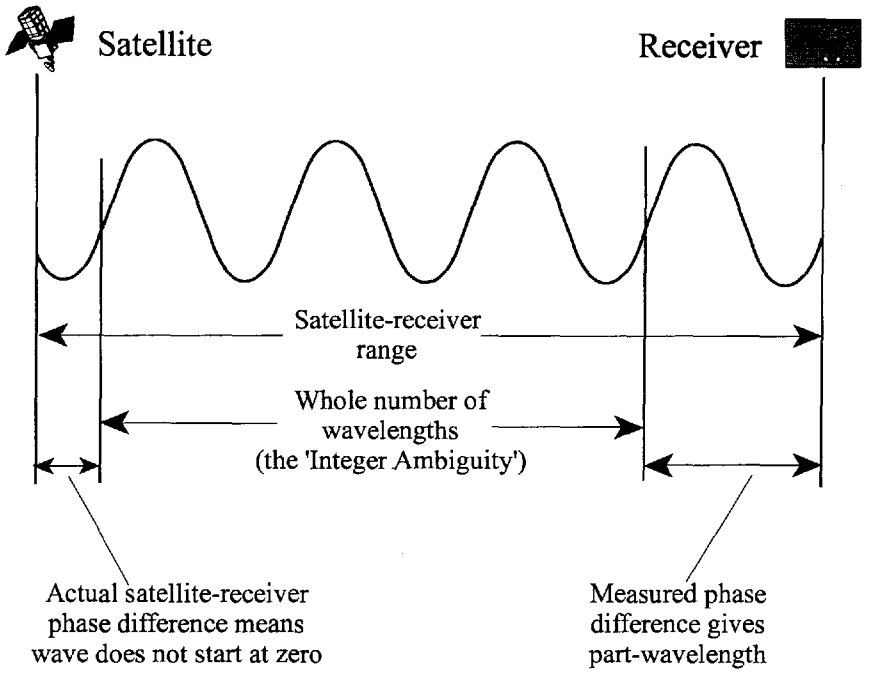

Figure 5.8 GPS phase observation.

as discussed in section 4.1.1) and improved corrections to the satellite clocks, it is possible to solve all the parameters, including the three-dimensional coordinates of the receiver. By definition this cannot be done in real time, and it requires very sophisticated software, but it is capable of determining the position of the point to a precision of around 2 cm. This type of observation is most appropriately employed in global geodynamic studies.

- Alternatively, less time-intensive methods of dealing with the situation are based on making the observations in differential mode, using two or more receivers. Part of the advantage of this technique comes from the fact that it is no longer necessary to determine the integer number of wavelengths from the satellite to the receiver, but only the difference in the number of wavelengths from a satellite to each receiver.

The real advantage of GPS, and what has made it a tool that has all but replaced conventional control surveys, comes from exploiting the integer nature of the ambiguities, thus avoiding the need to collect data over a long period of time. Essentially the approach is to gather just sufficient data for the processing software to be able to recognise the integer values of these numbers. Thus, for example, if the solution to the equations yielded as ambiguities the numbers:

451.999 43.002 875.998

then, barring coincidence, these would be recognised as 452, 43, and 876 respectively. The situation now is that the relative distances between satellites and receivers are in

principle known to the precision of the phase observations, a matter of a few millimetres. This would lead to a baseline between the stations that is precise to a similar amount. Conversely, if a mistake had been made and the ambiguities incorrectly identified, an error of several decimetres would be introduced.

Thus, the use of GPS over short baselines is centred around the length of time required to resolve the ambiguities to the correct values *with a high degree of confidence*. Advances in modelling software have reduced the time required to around 5–10 min over baselines up to around 10 km long, and 1–2 min over shorter baselines. The techniques involved are greatly assisted by the ability of modern receivers to read the phase of the second carrier wave, L2, albeit with an increased amount of noise: this is achieved in different ways by different makes of receiver, but most are based on a technique of cross-correlating the P codes between the two frequencies. Generally speaking, however, this second frequency is used only to acquire additional data for a quick solution and not to make any correction for ionospheric refraction: over short distances, the assumption is that this will cancel out between the two stations, which is the main factor that imposes a distance limit to this technique.

Over most distances less than around 10–20 km, this assumption is usually valid, and the main error source would be multipath: by definition this is site specific and does not cancel out between stations. For phase observations this can potentially lead to errors of up to 5 cm in the observation to any one satellite, and besides the obvious solution of good site selection (which is sometimes impractical) the main way around the problem is to increase the period of observation beyond the minimum required to resolve the ambiguities.

For longer baselines, the effects of ionospheric refraction are potentially more serious. For this reason, the emphasis is no longer on obtaining rapid solutions by identifying the integer ambiguities, but on correcting for the effects of refraction by exploiting both carrier frequencies. This leads to much longer occupation times, usually measured at least in hours, sometimes days.

In summary, differential phase observations with GPS are capable of determining baseline vectors between receivers to a precision of around 1 cm: the length of time required for this will vary from a few minutes to a matter of days, depending on the length of the vector.

A slightly lower level of accuracy can be achieved more quickly over shorter distances through the use of *kinematic* techniques. The basic premise of these is that, once the initial determination of the integer ambiguities has been made, these values will stay the same even when the roving receiver moves to a new location, provided that the receiver continuously tracks the signal from the satellite. As with differential GPS using code observations, it is possible to operate kinematic phase GPS in a real-time mode. This is usually referred to simply as *real-time kinematic* (RTK). The data link is usually provided by a dedicated radio link: in the future, if it becomes possible to use the system over longer lines, communication by mobile phones may be more appropriate.

In principle, the accuracy of kinematic GPS is similar to that of conventional phase observations in the static mode. It will always be slightly lower, however, as there is no averaging over time, which means that the effect of multipath goes uncorrected. Typically, at any one point there will be an error of around 2–3 cm.

The procedure for kinematic GPS is first to initialise the roving receiver by acquiring enough data to resolve the integer ambiguities. This can be done while the receiver is stationary; alternatively, with the 'on the fly' (OTF) techniques that have now been developed, data may be collected while the receiver is in motion. For OTF, the position of the receiver during the initialisation period can be deduced after the event, but not in real time.

Even if an individual signal is interrupted, it is possible to proceed provided that at least four satellites are tracked continuously. This is because four satellites is the minimum requirement for positioning and, if the position is known, when new satellites are observed or new ones re-acquired, the resolution of the ambiguities is instantaneous as the position is known. A complete blockage of all signals, such as would be caused by passing under a bridge, cannot be supported, however. Under these circumstances, it is necessary to go through the initialisation cycle once again: repeated interruptions such as would occur in a cluttered urban environment would cause problems.

The existence of a system capable of measuring to a precision of 1 cm over several kilometres in a short period of time has revolutionised many tasks in surveying and geodesy. It must be emphasised, however, that what is obtained is a three-dimensional vector in the WGS84 system. To exploit the accuracy of this to the full, great care must be taken in the way in which the results are incorporated into a local datum. This subject is rather more extensive in its philosophy and application than the simple transformations that have been discussed so far, and is therefore the subject of a more extensive treatment in section 6.5.

6
Aspects of datum transformations

6.1 Introduction

Chapters 2–4 have defined two- and three-dimensional coordinate systems, and discussed the procedure for transforming from one datum to another. If all the data and parameters mentioned there are known, there will not be a problem. It is often (perhaps usually) the case, however, that some or all of the information required is missing. The aim of this chapter is, firstly, to clarify where some of the problems are likely to occur. For these situations, the likely magnitudes of the errors caused is examined: often, missing information is not really a problem at all, as it has little effect at the level of accuracy that is required. If missing information is likely to cause a problem, various shortcuts are suggested.

In general, the approach taken will be very much dependent on the accuracy requirements. Since we are interested on the one hand with remote sensing and GIS applications that will usually be dealing with data accurate to at best 1 m, and on the other hand with surveying projects requiring an accuracy of perhaps 1 cm, the approaches differ so widely that it is most convenient to treat them separately. Sections 6.2–6.4 should therefore be read as a general introduction to some of the problems, but with the solutions proposed being relevant mainly to low accuracy applications. Section 6.5 deals with the high accuracy applications of GPS in surveying projects.

6.2 Knowledge of separation, N

It was stated in Chapter 1, with reference to Fig. 1.3, that a knowledge of the spheroidal height, and hence the separation, N, is necessary for two-dimensional coordinate transformations. This point is illustrated in Fig. 6.1, in which an error in the separation (perhaps equating it to zero) effectively shifts the point P to the point P′. In system A the point P′ has the same two-dimensional coordinates as point P, but in system B the

two points have different coordinates. The order of magnitude of this error is given by

$$\varepsilon = N\zeta \tag{6.1}$$

where ε is the error in the two-dimensional coordinates, latitude and longitude (in metres), N is the separation (in metres) and ζ is the angle between the two normals (in radians). Because N is at most 100 m, and a value for ζ of $20''$ (arc seconds) is quite large, the effect on (ϕ, λ) of ignoring the separation (effectively saying that $h = H$) is of the order of 1 cm. This is significant in some geodetic and surveying applications, but can safely be ignored in the context of remote sensing and GIS.

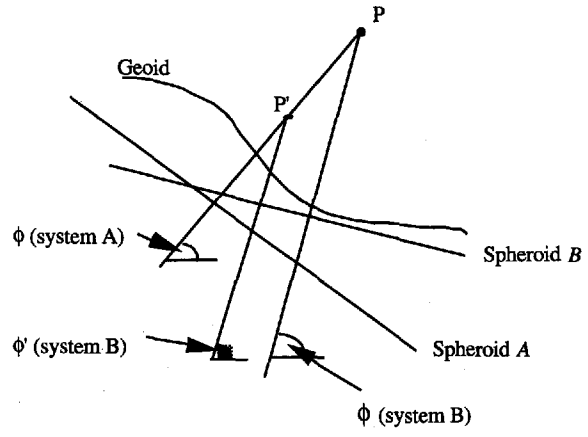

Figure 6.1 The effects of an error in separation.

6.3 Knowledge of height, H

In many situations it is necessary to transform a set of two-dimensional information for which no values of height are available (e.g. from a map without contours). A very similar problem to that in section 6.2 then results, but of a larger magnitude. Another look at Fig. 6.1 will show that if the height of the point P is assumed to be zero, a false point P' with different coordinates in system B results. In this situation, the order of magnitude of the error is

$$\varepsilon = H\zeta \tag{6.2}$$

where H is the height above the geoid. In this case, 1000 m represents a large value, if not a limit, and the order of the error is then seen to be around 10 cm.

It seems safe to assume that two-dimensional transformations to a precision of within 1 m can always be carried out without a knowledge of height.

6.4 Knowledge of datum transformation parameters

6.4.1 Sources of information

This is the most problematical aspect of the datum transformation procedure. For most datums the parameters of the spheroid are known. This is basically because all surveys have always had to use these values, even before the advent of satellite geodesy. In most cases a spheroid was taken 'off the shelf', as in the days before computers this saved a considerable amount of computation. It should be noted, however, that the parameters of a spheroid were occasionally modified, and it is therefore worth checking the actual values of a and f that have been used if this is possible.

A spheroid is often quoted by name, for example Airy or Clarke 1880. There is a danger, however, that 'modified Clarke 1880' is misquoted as 'Clarke 1880' (or, indeed, confused with Clarke 1880 (IGN), Clarke 1880 Palestine, Clarke 1880 I, Clarke 1880 II, or Clarke 1880 III).

The transformation parameters from a datum to WGS84 (ΔX, ΔY, ΔZ, plus rotations and scale if necessary) are more of a problem. In some cases these are regarded as classified information. In most developed countries the values are known (within the accuracy limitations explained in section 3.1), and are held by the relevant mapping authority. It is far more satisfactory, however, to have a central source of information to refer to. Several reports on this have been published by the Defense Mapping Agency (DMA, 1997), the latest addition including additional GPS survey information to update the datum transformation tables. For the European region, an alternative is University of Zurich (1989).

For high precision uses of GPS, the derivation of transformation parameters is described in section 6.5.

6.4.2 Estimating the parameters

If no information is forthcoming from any of the sources in section 6.4.1, one possible alternative is to attempt to estimate the transformation parameters from a knowledge of the geoid in the region. As a demonstration of this technique, consider the situation in the UK as an example.

The British datum, OSGB36, was effectively established by maintaining the position of the spheroid at Greenwich from the nineteenth-century survey. Therefore OSGB36 is parallel to and coincident with the geoid at the point:

$$\phi = 51^\circ\ 28'\ 40''\ \text{N} \qquad \lambda = 00^\circ\ 00'\ 00''\ \text{E}$$

At this point the geoid–spheroid separation, N, and the deviation of the vertical, ζ, are both zero in OSGB36. Using a global Earth model such as EGM96, the values of N and the components of ζ can be computed with respect to the WGS84 datum. The

relevant values are:

$N = 45.796$ (separation)
$\xi = -2''$ (component in latitude)
$\eta = +3''$ (component in longitude)

Using these values, the coordinates of the point at Greenwich can now be found in both the local system and WGS84. Adopting the terminology of Fig. 1.3 and converting both sets of coordinates to cartesian values gives

- for WGS84:

$$\begin{array}{|lll|} \hline \phi & = & 51^\circ\,28'\,42''\,\mathrm{N} \\ \lambda & = & 00^\circ\,00'\,03''\,\mathrm{W} \\ h & = & N = 45.796 \\ \hline \end{array} \longrightarrow \begin{array}{|lll|} \hline X & = & 3980611 \\ Y & = & 0 \\ Z & = & 4966859 \\ \hline \end{array}$$

- for OSGB36:

$$\begin{array}{|lll|} \hline \phi & = & 51^\circ\,28'\,40''\,\mathrm{N} \\ \lambda & = & 00^\circ\,00'\,00''\,\mathrm{W} \\ h & = & N = 0 \\ \hline \end{array} \longrightarrow \begin{array}{|lll|} \hline X & = & 3980244 \\ Y & = & 58 \\ Z & = & 4966420 \\ \hline \end{array}$$

A direct comparison of the cartesian coordinates then yields the transformation parameters. Putting them alongside the known values gives the results shown in Table 6.1. The method has given very good results in the X and Z directions (within 10 m), but a value for Y that is over 50 m from the known value. The most probable cause of this is a slight rotation between the two datums, which on the Greenwich meridian would manifest itself as a discrepancy in the Y direction. This geometrical approach is not able to detect this effect, and rotation errors of this magnitude are likely to result from any similar attempt to estimate the transformation parameters.

Table 6.1 Attempts to derive transformation parameters

	ΔX	ΔY	ΔZ
True values	−375	111	−431
Estimates	−367	58	−439

Another potential source of error when applying this technique is that the point of origin is likely to be unknown. Therefore, attempting to estimate the transformation parameters from such a geometrical approach should be used only for datums that cover a very limited area, and when there is no possible alternative.

6.4.3 Simple two-dimensional transformations

In some cases it may be necessary to transform two-dimensional data on one datum into another datum. The transformation parameters may be unknown, or a simpler

transformation procedure than the one outlined in Fig. 1.3 may be required. An example of this might be maps of adjacent countries which are on different datums.

If common points can be identified in both systems, then one possibility is to transform from one datum to the other by a two-dimensional transformation, either a simple similarity transformation or a more complex affine or polynomial one (these are treated more extensively in Chapter 12). The extent to which this is possible will depend on the extent to which the relative positions of points are altered in transforming from one datum to another.

As an example of the effect of a datum transformation on coordinates, consider the test area of approximately 100×100 km that is defined by the four points in Table 6.2. The coordinates of Table 6.2 are defined on a datum with the spheroidal parameters

$$a = 6378388 \text{ m} \qquad f = 1/297$$

These are converted to a datum with the spheroidal parameters

$$a = 6378136 \text{ m} \qquad f = 1/298.257$$

The following datum transformation parameters are applied:

$$\Delta X = 200 \text{ m} \qquad \Delta Y = 200 \text{ m} \qquad \Delta Z = 200 \text{ m}$$

Table 6.2 Corner points of 100 km test square

Point	Latitude	Longitude
A	50° 00′	00° 00′
B	50° 00′	01° 24′
C	50° 54′	00° 00′
D	50° 54′	01° 24′

The resulting coordinates on the new datum are shown in Table 6.3. Converting these into *coordinate changes* in metres gives the results shown in Table 6.4. Although the changes are substantial, it can be seen that the spread is less than 8 m: a similarity transformation could therefore model this effect to within a couple of metres. This is not accurate enough for geodetic or surveying applications, but in many other contexts it is more than sufficient.

Table 6.3 Coordinates on the new datum

Point	Latitude	Longitude
A	49° 59′ 56.29″	00°00′10.04″
B	49° 59′ 56.17″	01°24′09.79″
C	50° 53′ 56.16″	00°00′10.23″
D	50° 53′ 56.04″	01°24′09.98″

In practice, the most common source of problems in applying a two-dimensional transformation to convert between data sets is the change in the projection rather than the change in datum; for this reason, the subject is given a somewhat more extensive treatment in Chapter 12.

Table 6.4 Coordinate changes after transformation

Point	Latitude	Longitude
A	−115.0485	200.09047
B	−118.7578	195.14186
C	−118.9318	200.07987
D	−122.689	195.13152

6.5 Datum transformations for precise applications

6.5.1 Distorting to fit the local datum – or not, as the case may be

This section considers the datum problems that arise when dealing with applications that require a high level of accuracy when transforming from one datum to another. The most obvious example is in surveying points with GPS using differential phase observations, and when the final coordinates are required to be expressed in a local datum such as OSGB36.

To begin with, it is helpful to consider what the final aim of the survey is; that is, the use to be made of the coordinate information that is derived. This has a particular bearing in situations where GPS data is being transformed into a datum that contains the types of distortion discussed in section 4.2.2.

The problem is essentially one of how closely these distortions should be followed when transforming. To take an example, an engineering project (such as the construction of a road or railway) should usually use coordinates that are in sympathy with a local coordinate system. On the other hand, building a structure in the right place from the point of view of, say, land ownership, is not a very exacting requirement (20–50 cm might suffice, for example). By comparison, it is quite important that the internal geometry of the project should be preserved: it is not acceptable to have kinks in the railway where the local datum is distorted, and it is usually helpful if measurements carried out with ground survey equipment can build onto the GPS control without a continuous need for distortions to be introduced. To take a counterexample, if GPS is being used to locate features whose coordinates are given in the local system (underground utilities, perhaps), transforming precisely into the local system – with all its distortions – is of paramount importance.

One way of treating either of these situations is to derive the transformation parameters as part of the project, by using points with coordinates that are known in the local system (for example Ordnance Survey triangulation points). For situations where the internal geometry of the project is of more importance than sympathy with existing mapping, relatively few control points are needed. For the opposite case, a large number of control points, well distributed across the survey area, will be necessary. In the latter case, an alternative is to use a transformation service provided by an organisation such as the Ordnance Survey. The advantage of this is a consistency between the transformations used by different organisations; the disadvantages are that it is not free and it is less appropriate for engineering projects that require consistent geometry.

One of the main problems that will be encountered in deriving coordinates in the

local system is the influence of the geoid. Moreover, the procedure is not uniquely concerned with datums: a knowledge of the projection used for the local coordinates will also be required. It will be assumed in this section that all information pertaining to the projection is available: readers unfamiliar with map projections should refer to Chapter 7.

6.5.2 Transformation models

In order to derive transformation parameters between WGS84 and a local system, it is necessary to include several known points in the survey. An example of such a scheme is shown in Fig. 6.2. The first point to be noted about Fig. 6.2 is that the known points are well distributed across the area to be surveyed, rather than being grouped to one side. This is important, as the parameters derived will be applicable only in the area of the known control points and extrapolation beyond this area is likely to cause problems. An example of an unacceptable configuration is shown in Fig. 6.3.

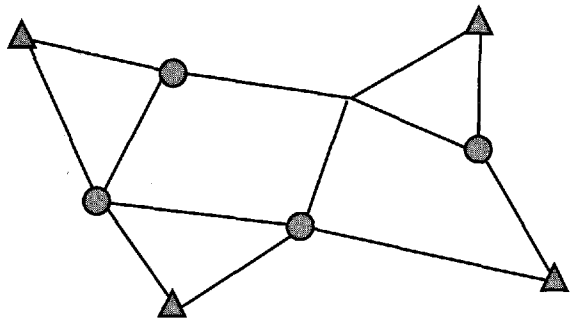

Figure 6.2 Including known points in a survey. Triangles indicate points known in the local system; circles are new points; the lines denote GPS vectors.

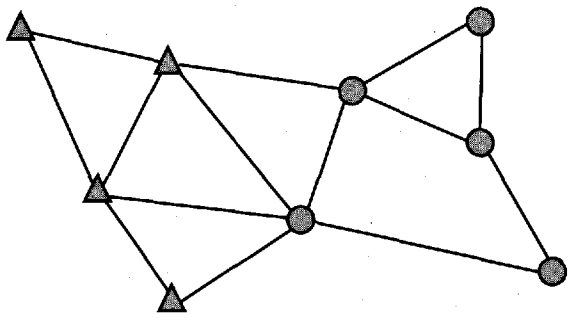

Figure 6.3 Poor geometry of known and unknown points.

For very small surveys, it is possible to use only one known point for deriving the transformation, but this would necessarily involve using the very simple translation given in equation (5.1) and ignoring the rotations. As the latter would typically be around $10''$ (due to the combined effect of coordinate errors and the geoid), this

approach would be valid only over distances up to 200 m if an accuracy of 1 cm is required. The rest of this section will assume that more extensive surveys are under consideration.

It is a requirement of using GPS in the relative mode that the absolute coordinates in WGS84 are known to an acceptable level of accuracy. For surveys that aim to achieve a relative accuracy of 1 cm over 10–20 km, the absolute coordinates should be known to within 10 m. If the survey has been designed to include a point that is part of a more accurate control survey, this should not be a problem. Alternatively, as outlined in section 5.3, absolute coordinates of this accuracy can be obtained with around 8 h of observation at a single point. The result is then a set of coordinates that are accurately known with respect to one another, but 'floating' in WGS84 by up to 10 m. Again, the concept of a 'quasi-WGS84' datum is useful here. It is necessary that the *same* 'quasi-WGS84' datum is used throughout the project, and therefore that either the absolute coordinates of only one point are determined, or a 'best fit' is carried out to the absolute determination of position made at all points in the survey.

The task is then to transform from this datum to the local one. A possible model for the transformation equations was shown in section 4.2.2 and equations (4.3)–(4.4). This is not the only possibility, however. An alternative model is one that makes the rotation about a point at the centre of the points to be transformed, rather than the centre of the coordinate system. The result will be the same, but the latter approach has the advantage of making the transformation parameters so derived easier to interpret. In the former approach, a rotation around the coordinate origin (at the centre of the spheroid, several thousand kilometres away) is very similar to a translation, and a large additional shift may be needed to compensate for this. Where the rotation is about a more local point, there is very little correlation between the shifts and the rotations, and the shifts derived will be much closer to the typical value for the offset between the two datums.

The transformation about a local origin is expressed through:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{\text{WGS84}} = \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} + \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix} + \mu \mathbf{R} \begin{pmatrix} X - X_0 \\ Y - Y_0 \\ Z - Z_0 \end{pmatrix}_{\text{local}} \tag{6.3}$$

where

$$\begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix}$$

are the local coordinates of a point at the centre of the survey, and other terms are as defined in section 4.2.2. Note that both of these models are referred to as *similarity transformations*, as they preserve the shape of the original data.

The task now is to derive the seven unknown parameters (three translations, three rotations, and one scale) of the transformation model by comparing the coordinates from the two systems. Any commercial GPS package will be able to do this: the least squares formulation is given in Appendix section A3. It should also be noted that this procedure requires both sets of data to be in cartesian form, and therefore

coordinates quoted in a projected coordinate system and with orthometric heights must first be transformed to the cartesian form (this is a standard part of most GPS software packages).

6.5.3 The use of redundant data

To derive the seven parameters of a similarity transformation, seven pieces of information are needed. This could in theory be provided by two control points that are known in three dimensions and one bench mark (known in height only). In fact, the more control points that are available the better. Once the parameters have been derived, it is possible to apply these to the known points and note the agreement between the original coordinates and the transformed values. If there is no redundancy the fit will be perfect, even if there was an error in the coordinates given for the known points.

Such an error could occur through incorrect keying-in of the data; alternative sources of error are that the control point could have moved (for example through local subsidence) since the last publication of its coordinates, or even that a monument may have been demolished and rebuilt in a slightly different location. In effect, this would mean that, in an effort to fit the GPS data to the distortions of the local system, *additional* distortions have been introduced that are not justified.

Therefore, a minimum of three control points, each known in three dimensions, is needed in order to provide redundancy and check for errors. Even in this case, with nine pieces of information, the redundancy is only 2. There would be a tendency for the residuals (the difference between the original coordinates and the transformed values) to be much smaller than the size of any errors present: the transformation would stretch things a bit, rotate a bit more, and generally do all it could to fit to the given values. An example of this is given in the case study in section 13.1.

Therefore, it must be emphasised once again that as many control points as possible should be used, and that they should be well distributed across the survey area.

6.5.4 Geoid problems

The classic procedure for datum transformations outlined in section 1.5 assumes that the geoid separation is known. If it is, it should be added to the orthometric heights quoted for the control points to derive heights above the spheroid.

This section will consider the implications of geoid information *not* being available. By way of illustration, a hypothetical data set will be used with different geoid characteristics. The points used are shown in Fig. 6.4. In the figure, points A, B, C, and D are known. The coordinates of P are also known (see Table 6.5), but will be treated as unknown in order to test the quality of the transformation.

The four known points cover an area of 25 km square. It is assumed that the local datum is parallel to WGS84, but translated by 100 m in each dimension, and that there are no distortions present in the local datum. Such a paragon of perfection is unlikely, and is used here only to isolate the effect of the geoid. For a more realistic example, in which several sources of error are present simultaneously, see the case study in section 13.1.

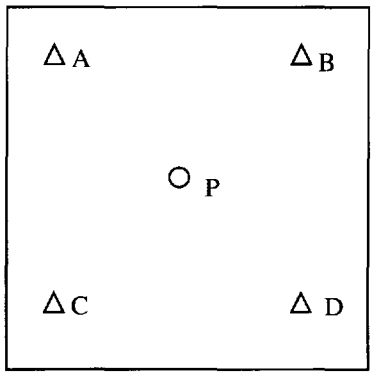

Figure 6.4 Test data set.

The first case to consider is the one in which there is a uniform separation of the geoid across the whole area. If the geoid separation is unknown, the only option is to enter it as zero and assume that the ellipsoidal heights are the same as the orthometric heights. In this case, the coordinates of the control points will all be shifted. For a sufficiently small area, however, they will all be shifted by the same amount (for a uniform value of the separation) and in the same direction. This is illustrated in Fig. 6.5.

A seven-parameter similarity transformation can cope with a uniform shift, as this will be included in the translations that are derived. What does cause a problem is a change of shape of the control points. The extent to which this happens is a function of the maximum angle between the shifts caused at the different control points, which is the same as the angle between the spheroidal normals. A general rule of thumb for the relative coordinate shift caused by this effect is

$$\varepsilon = \frac{D}{6400} N \tag{6.4}$$

where ε is the resulting error, D is the extent of the survey (in kilometres), and N is the approximate size of the separation that has been ignored (and has the same units as ε). For a survey of 25 km in extent and a separation of 10 m, this amounts to around 0.04 m. This is significant at the required accuracy level, although a certain amount of this effect is absorbed by the scale factor of the transformation.

To illustrate this, let us assume that the geoid separation is a uniform 10 m in the test area. Thus, although the heights are all 50 m above the ellipsoid, they will

Table 6.5 Coordinates of test points

Point	E	N	h
A	400 000.00	125 000.00	50.000
B	425 000.00	125 000.00	50.000
C	400 000.00	100 000.00	50.000
D	425 000.00	100 000.00	50.000
P	412 500.00	112 500.00	50.000

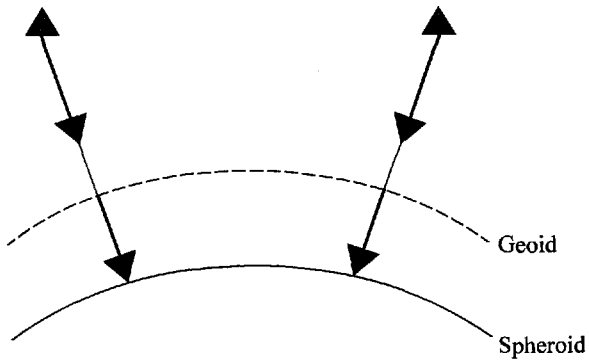

Figure 6.5 A uniform value of the separation.

Table 6.6 Transformation parameters for uniform geoid shift

ΔX	−106.302 m
ΔY	−99.799 m
ΔZ	−107.762 m
μ	−1.57 ppm

actually be quoted as 40 m above the geoid. If the transformation algorithm then assumes (incorrectly) that these are spheroidal heights, the transformation parameters are as shown in Table 6.6. The values of the rotations are not shown in the table, as they are all nearly zero.

Clearly these parameters are 'wrong' in the sense that the translation is known to be 100 m in each dimension, and the scale should be true. What has happened, however, is that the extra 10 m shift caused by the geoid has been absorbed in the transformation.

Applying these parameters to the whole data set results in coordinates for P as shown in Table 6.7. In other words, the plan coordinates are correct to a millimetre and the spheroidal height that the program thinks it has found for P is the correct geoid height.

As well as being able to cope with a uniform shift of the geoid, a similarity transformation can also deal with a uniform tilt, by adjusting the rotations determined during the transformation. This can be illustrated in the test data set by assuming that the geoid slopes as a plane from a height of 10 m on the western side (in line with A and C) to 11 m on the eastern side (in line with B and D). Therefore the levelled heights of the four control points will be as shown in Table 6.8.

Applying the same procedure as before results in the transformation parameters shown in Table 6.9. Although the other parameters are much the same as before,

Table 6.7 Coordinates of point P after transformation

Point	E	N	h
P	412 500.000	112 500.000	40.000

Table 6.8 Orthometric heights for uniform slope of the geoid

Point	H
A	40.000
B	39.000
C	40.000
D	39.000

Table 6.9 Transformation parameters for uniform geoid tilt

ΔX	−106.617 m
ΔY	−99.790 m
ΔZ	−108.150 m
Rotation of X	$6.40''$
Rotation of Y	$-0.22''$
Rotation of Z	$-5.20''$
μ	−1.65 ppm

the rotations are now significantly different, as the coordinates rotate to adapt to the sloping geoid.

Applying these parameters to the data set gives the results shown in Table 6.10. The plan coordinates are again correct within 1 mm, and the height is also the expected value, as a uniform slope of the geoid would result in a separation of 10.5 m at P.

The similarity transformation is thus very adept at dealing with uniform slopes and shifts of the geoid. This will quite often suffice for surveys that cover just a few kilometres and are not seeking the highest accuracy. What the similarity transformation *cannot* cope with is a non-uniform change in the geoid: perhaps a bulge or smaller undulations. In the previous example, as the point P was not included in the derivation of the transformation parameters, any value of the separation other than that implied by a uniform tilt and shift of the geoid would not have been corrected by this procedure.

If the undulations of the geoid are sufficiently large to cause a problem at the accuracy required, and no model of the geoid is available, the only solution to this problem is to incorporate more information into the transformation and to alter the model used. It is possible, for example, that, although no extra triangulation points are available in the area surveyed, there may be several bench marks. If these points are included in the GPS survey, a comparison may be made of the heights derived by the similarity transformation with the original bench mark heights. The differences between these two figures will not represent the geoid–spheroid separation itself, but in the absence of observational errors will represent a residual value of the separation over and above the overall shift and tilt.

If the residuals so found are indeed due to finer undulations of the geoid, the expectation would be that they are highly correlated, certainly over short distances.

Table 6.10 Coordinates of point P after transformation

Point	E	N	h
P	412 500.000	112 500.000	39.500

A random spread of residuals with no apparent pattern is an indication more of poor quality GPS data or bench mark heights than of short wavelength undulations of the geoid (this is particularly the case in low-lying terrain). Where the residuals do display a pattern it is possible to interpolate the geoid values between bench marks to obtain orthometric heights at the points newly surveyed by GPS. This could be done either by simple methods of interpolation, or by more sophisticated statistical techniques such as least squares collocation. A full description of the latter may be found in Moritz (1980), and an example of its application is given as part of the case study in section 13.1.

7
Fundamentals of map projections

7.1 Introduction

The fundamental coordinate system for surveying and mapping is a set of geodetic coordinates related to a particular datum. It is then necessary to consider how to arrange the data so that it can be placed on a flat surface. There are two reasons for doing this. The first, and most obvious, is presentational. Whether the data is to be shown on a paper map or on a computer screen, it must of necessity be presented in a two-dimensional format. The second reason for rearranging the geodetic coordinates in two dimensions is computational. Even a simple concept such as the distance between two points becomes excessively complex when expressed in spheroidal formulae, and wherever possible it is more desirable to carry out computations in a simple two-dimensional coordinate system. A projection, then, is defined as an ordered system of meridians and parallels on a flat surface. It should be immediately apparent that it is impossible to convert a sphere or a spheroid into a flat plane without in some way distorting or cutting it. It follows that there is no single method for doing this; hence the proliferation of types of map projection.

This chapter is concerned with introducing some of the fundamental concepts of map projections, before looking in detail at the different types of projection. Included here are definitions of grids and graticules, and of scale factor, as well as a consideration of the use of spherical or spheroidal models of the Earth in the context of projections, the use of different developable surfaces, and the criteria to be considered in designing a projection for a specific purpose. Some fundamental defining parameters have more conveniently been introduced in the context of specific projections, although they have a universal application. For reference, these are:

- *false coordinates* and the *projection origin*: discussed in section 8.1 with reference to the cylindrical equidistant projection
- *re-scaling* of projections: discussed in section 8.3 with reference to the Mercator projection

- *zoning*: discussed in section 8.4 with reference to the transverse Mercator projection
- *convergence*, or the angle between grid north and true north: also introduced in section 8.4.

An example of the development of formulae to convert between geodetic and projection coordinates is given in section 8.1. A summary of the parameters needed to define a projection is given in Chapter 11.

Chapters 8–10 consider many different methods for projecting coordinates: it will be seen that each of these methods requires several defining parameters. It is therefore important to distinguish between a *projection method* (for example the transverse Mercator, the polar stereographic, and so on) and a *projected coordinate system*, which is composed of the method as well as the defining parameters (for example the British National Grid, the Universal Transverse Mercator system, and so on). This distinction in terminology is a useful one, and will be encountered, for example, in the definitions of the GeoTIFF format for transferring geographically referenced data (GeoTIFF, 1995).

7.2 Spheres and spheroids

As mentioned in the introduction, the fundamental coordinate system is a geodetic one related to a spheroid. The relative positions of points in such a coordinate system are not the same as they would be on a sphere. That is to say, if accuracy is to be preserved it is necessary to develop formulae for treating a spheroid rather than a sphere. That said, it should be borne in mind that the flattening of most spheroids is of the order of 1 part in 300. It is therefore apparent that the shape of a particular projection when applied to the sphere is very similar to the shape when it is applied to the spheroid. There are significant differences in the coordinates that result, which will certainly be apparent on a large scale map, but the sphere is nevertheless useful for giving an insight into how the resulting map has been distorted. Most of the explanation that follows therefore uses the sphere as a model of the Earth. Full spheroidal formulae should, however, normally be used in practice.

7.3 Grids and graticules

The appearance of meridians and parallels on the projection depends on the type of projection that has been used. In general they are an arrangement of straight or curved lines, as shown for example in Fig. 7.1. This set of parallels and meridians, as seen on the map, is known as the *graticule*. On some maps it may not be shown at all; on others it may be noted around the border of the map and shown in the middle as a set of tick marks (for example on Ordnance Survey 1 : 50 000 maps, where blue tick marks show the graticule at $5'$ intervals).

This graticule does not constitute the basis of a coordinate system that is suitable for computational purposes or for placing features on the projection. Instead, a rectangular coordinate system known as the *grid* is superimposed on the map (see Fig. 7.2).

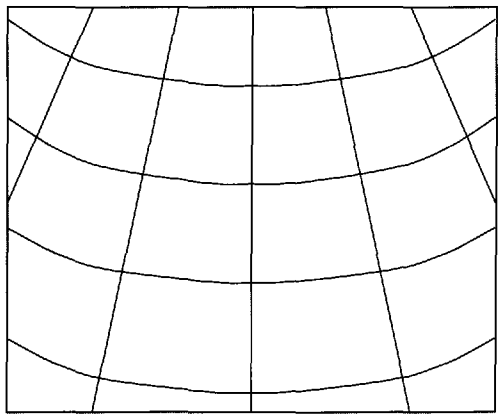

Figure 7.1 An example of the graticule.

This grid may be given coordinates x and y, or if appropriate eastings (E) and northings (N). This may seem a rather obvious point, but it is important to establish the difference at this stage between a graticule and a grid.

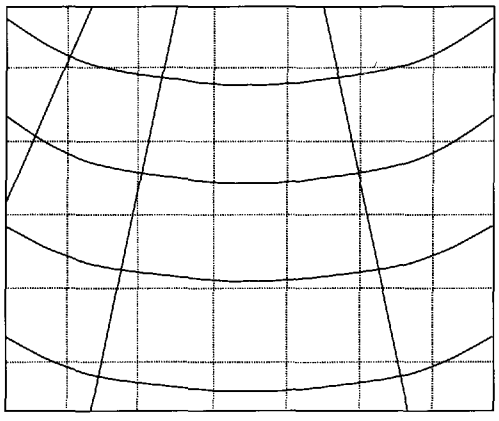

Figure 7.2 A grid superimposed on the graticule.

7.4 Scale factor

Features on the surface of a sphere or a spheroid undergo distortions when projected onto a plane. It is necessary to have a precise definition of the amount of distortion that has resulted. This is provided by the definition of the *scale factor*, which is given the symbol k in this text. Then

$$k = \frac{\text{distance on the projection}}{\text{distance on the sphere}} \tag{7.1}$$

This parameter will be different at each point on the projection, and in many cases will have different values in each direction. Equation (7.1) can therefore be understood to

apply in general to only a short distance (in theory infinitesimally short). For longer lines, the relevant parameter is the integrated mean of the point scale factors along the whole length of the line: section 7.7 discusses the point at which a short line becomes a long one.

It is important to understand that this scale factor results purely from the act of projecting to a flat surface and is therefore unrelated to the map scale (a number such as 1 : 50 000). The ideal value of scale factor is 1, representing no distortion. It should also be emphasised that a distortion of this type is not the same as an 'error' in the map, as the rules governing it are clearly defined and the true coordinates can always be recovered if the parameters of the projection are known.

It will often be useful to consider what happens to a small square of dimension (1×1) on the surface of the sphere when it is projected. In the general case the distortion in the direction of the parallels will be different from the distortion in the direction of the meridians. Let k_P represent the scale factor along a parallel, and k_M represent the scale factor along a meridian. With reference to Fig. 7.3, the square is then projected as a rectangle of dimensions $(k_P \times k_M)$. It will be seen in the examples of projections that follow that the unit square is often subjected to a rotation as well.

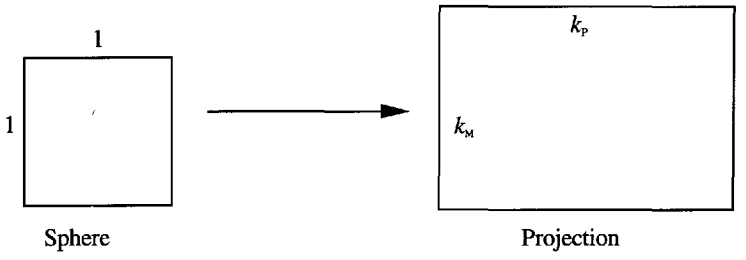

Figure 7.3 Projection of a unit square.

7.5 Developable surfaces

It is possible to derive a set of formulae to convert geodetic coordinates to grid coordinates in purely mathematical terms. Historically, projections were derived by first projecting from the sphere to an intermediate shape, which was of such a nature that it could be unravelled without distortion. This remains a useful concept for categorising and describing map projections. The principal forms of these intermediate surfaces are the cone, the cylinder and the plane itself. The advantage of these shapes is that, because their curvature is in one dimension only, they can be unravelled to a plane without any further distortion. The developable surface is brought into contact with the sphere, and a set of rules is formulated for the way in which features are taken from the sphere onto that surface. Examples are shown in Fig. 7.4.

The rules for transferring features from the sphere to the projection are discussed in section 7.6. Before looking at these in detail, however, the general point can be made that in the region around the point or line of contact between the two surfaces the scale factor distortion will be minimal. In fact, where the two surfaces are touching the scale factor will be equal to 1.

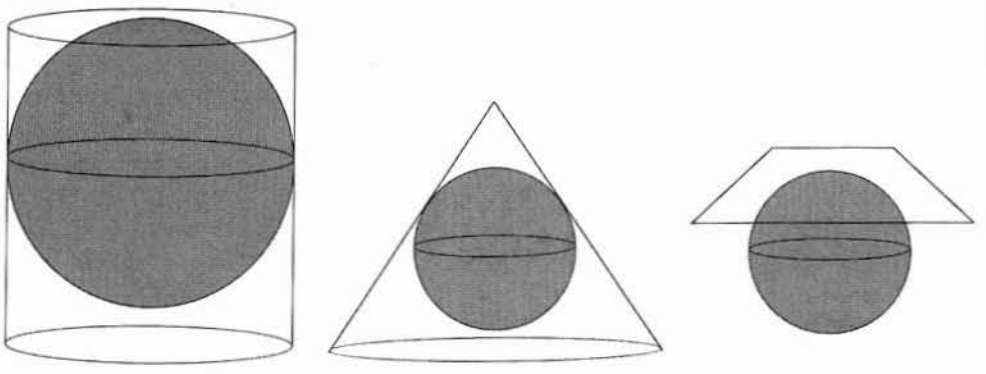

Figure 7.4 (a) cylindrical surface; (b) conic surface; (c) plane surface.

The choice of developable surface will therefore be dictated by the geographical extent of the region to be mapped. In some cases it is required to project the whole Earth; in others, however, a projection will apply only to a selected area. Thus, for example, a cylindrical surface (Fig. 7.4a) is appropriate for mapping the equatorial regions as it is touching the sphere along the equator. A conic projection (Fig. 7.4b) is good for mapping areas of mid-latitude with a large extent in longitude, as the cone is in contact with the sphere along a line of latitude.

The characteristics of the developable surfaces have only been given in outline here: it is, for example, possible to orientate the surface at different angles from those shown here, or to change the shape of the cone. These aspects are dealt with more fully in Chapters 8–10.

Cones, cylinders and planes are useful for gaining an insight into the appearance of a projection. This is extremely useful in situations where the parameters of the projection are unknown and it is necessary to try to guess them, a situation discussed in section 12.4. It should be noted here, however, that they are not a necessary step in forming a projection. In general, equations can be derived of the form:

$$(E,N) = f(\phi,\lambda) \tag{7.2}$$

which express the grid coordinates as a function of the geodetic coordinates without reference to intermediate developable surfaces. Indeed, the equations for all the above projections could be given in this form without mention of cones and so forth, and in this case expressions for aspects such as scale factor and convergence could be derived by differentiation.

Most of the more complex projections which depart from the simple forms in the following chapters are usually developed to represent the whole Earth in some way, and so are of less relevance to surveying, remote sensing and GIS. There are exceptions to this, however, such as the New Zealand map grid.

Most large scale mapping is likely to be based on transverse Mercator, Lambert conformal conic, and to a lesser extent the azimuthal stereographic and oblique Mercator projections.

7.6 Preserved features

Having selected the developable surface, it is necessary to devise a set of rules for transferring coordinates from the sphere. In theory, there is an infinite number of ways of doing this, and the choice will depend on the purpose for which the projection is devised.

It is not possible to devise a projection without introducing distortions. In general, the shape, area and size of features on the surface of the sphere will be different when transformed to the projection. The usual approach is to attempt to preserve *one* of these, usually at the expense of all the others. For example, it may be required that certain of the distances as measured on the sphere should be undistorted when shown on the projection. It is obviously not possible to preserve all distances, as this would then be achieving the unachievable goal of an undistorted projection. It may be instead that the distances along all meridians should remain undistorted, which is the same as saying that

$$k_M = 1 \tag{7.3}$$

or the scale factor along a meridian is equal to 1. The effect on the projection of a unit square is shown in Fig. 7.5. Such a projection is said to be *equidistant.* It can be seen that there remains a scale factor along the parallel which is not equal to 1, and that the shape and area of the square have both been distorted.

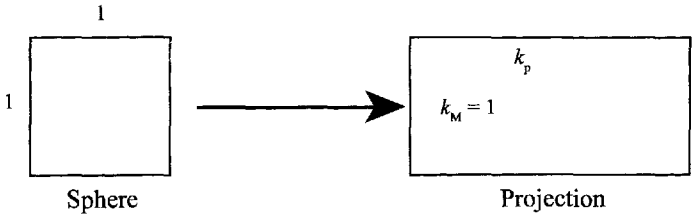

Figure 7.5 Projection of a unit square preserving distances along the meridians.

An alternative to this type of projection is one that attempts to preserve area, and is therefore termed an *equal area projection.* In such a situation we have

$$k_M k_P = 1, \tag{7.4}$$

or in other words the area of the projected unit square remains equal to 1. This is illustrated in Fig. 7.6.

The other principal classification of projections is that which preserves the *shape* of features. This is known as an *orthomorphic* or, more commonly, *conformal* projection, and the relationship between the scale factors is

$$k_M = k_P \tag{7.5}$$

This is illustrated in Fig. 7.7.

In preserving shape, a conformal projection is therefore preserving *angles* as well. For example, the angle between the side of the unit square and the diagonal is 45°:

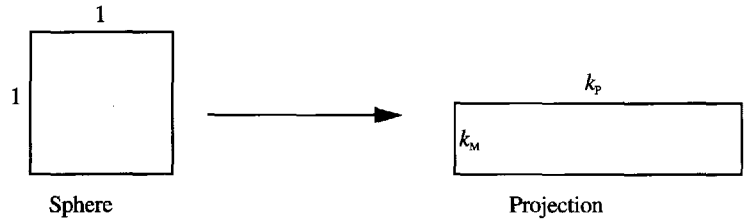

Figure 7.6 Projection of a unit square preserving area.

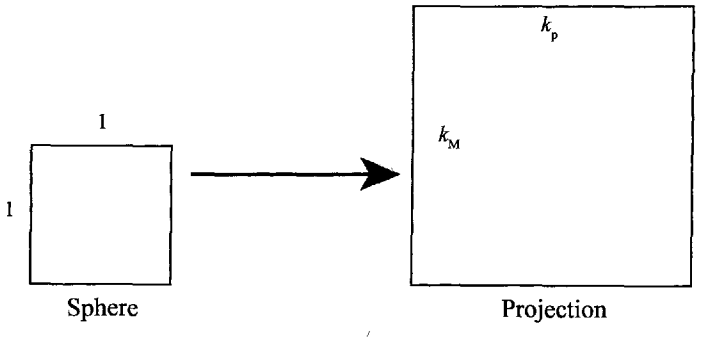

Figure 7.7 Projection of a unit square preserving shape.

this is the angle that would be measured by someone making the observation on the ground. In Fig. 7.7, the angle between the side and the diagonal is also 45°. This is not the case, however, in either Fig. 7.6 (the equal area projection) or Fig. 7.5 (the equidistant projection). For this reason, the conformal projection is the one of most significance in land surveying, as it means that angles measured on the ground can be transferred to the projection for use in computations. A conformal projection is therefore one of the most frequently used, and is likely to be the basis of almost all large scale mapping.

Finally, it should be noted that the three types of projection mentioned above, although the most commonly used of all projections, do not constitute an exhaustive list. Other types are possible, and are sometimes used, which preserve neither shape, area nor any distances. Some of these will be referred to in later sections.

It should be emphasised here that the unit square used in these examples has to be a very small one for the conclusions drawn here to be exact: for bodies of finite size, some of these assumptions can break down. This point is treated more extensively in Chapter 12.

In summary, most projections may be classified firstly according to the shape of the developable surface, which is dictated primarily by the geographical area to be mapped, but also in part by the function of the map, and secondly by the features on the sphere which are to be preserved on the projection.

7.7 Computational aspects

From a traditional land surveying point of view, there are two principal aspects to computing coordinates on projections. The first concerns distances, and the second concerns angles.

The basic principle of transferring a distance measured on the Earth to one to be used on a projection has essentially been covered by equation (7.1) in section 7.4. This can be rearranged as

$$\text{distance on the projection} = k \times \text{distance on the sphere} \tag{7.6}$$

That is, any measured distance must be multiplied by the appropriate scale factor in order to use it on the projection. This procedure is simplified somewhat by the fact that most projections used for survey computations are conformal. Hence, the scale factor at a point will be the same in all directions.

In section 7.4 it was pointed out that the definition of scale factor applies in theory to lines of infinitesimally short length. In practice, the scale factor changes so slowly across a projection that a single scale factor can often be considered as applicable to all the distances in one survey, rather than having to compute a separate one for each. How slowly? And how significant is the change in distance as a result of applying the scale factor? In part, the answers to these questions will depend on the particular projection but, as the area that any one survey projection covers is restricted in order to avoid excessive corrections of this type, a transverse Mercator projection applicable to a zone 6° wide may be taken as a suitable example and inferred as being typical of most survey projections. (Note the distinction here between a 'survey' projection used for base mapping and computational purposes and 'other' projections where the object may be the presentation of data and the distortions very large.)

A mid-latitude country mapped on a transverse Mercator projection has a maximum scale factor of around 1.0004. This means that a distance of 100 m as measured on the ground should be scaled to 100.04 m when used on the projection, a correction that is certainly significant when compared with the accuracy that may be obtained with an electromagnetic distance measurer (EDM).

The region with the most extreme rate of change of scale factor is on the edge of the projection. Over a distance of 5 km, the scale factor may, for example, vary from 1.00043 at one end of a line to 1.00039 at the other. The error introduced by using the scale factor at one end of the line, rather than the average computed over its whole length, would in this situation be around 8 cm. A similar calculation over a distance of 3 km in the worst case scenario shows a potential error of 3 cm.

For a target accuracy of 1 cm, and distances over 1 km, it would therefore be advisable to calculate a scale factor for each line. This can be done either by using the mean of the end point values of scale factor or by using Simpson's rule (Allan, 1997a) for higher precision. For a project spread over an area of less than 1 km^2, it will nearly always be acceptable for a general scale factor to be used for the project as a whole.

The correction to be applied to angles arises because, in general, the straight line observed between two points does not plot as a straight line on the projection. Standard trigonometric calculations on the coordinates, distances, and angles assume that all lines are straight, and so a correction has to be applied to obtain the angle that

would have been observed along the straight lines on the projection. This is illustrated in Fig. 7.8. As the magnitude of this correction is proportional to the length of the line, and is never large, its significance has greatly diminished with the use of GPS in place of triangulation over long distances. At its most extreme on the UK National Grid it reaches 3″ over a distance of 5 km, and 6″ over 10 km. For the vast majority of modern surveys it may therefore be safely dismissed. Otherwise, a text such as Allan (1997b) may be consulted.

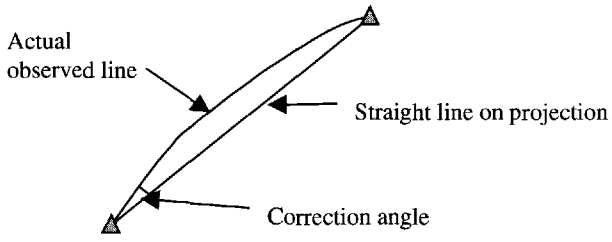

Figure 7.8 Difference between actual line and projection line.

7.8 Designing a projection

In many applications, those working with map projections will be using an existing projected coordinate system, and the task will be to identify the projection method and the associated parameters. In other situations it is necessary to design a projection for a particular purpose, in which case the choice of projection method and parameters is up to the user. In such a case it is first necessary to define what is meant by a suitable projection. A suggested order of criteria is given below.

1. It should preserve any properties that the use of the map dictates. That is, if the map is to be used to measure the areas of features, the projection must be an equal area projection. Alternatively, if it is to be used for surveying or navigation, the shape must be preserved and the projection has to be conformal.

2. A good projection is one that minimises the scale factor over the region; that is, the scale factor must be everywhere as close to unity as possible. In doing this there may be some goal for the maximum allowable scale factor distortion, which may lead to a situation where a single projection cannot achieve the desired result and the area must be split up into zones. This adds to the complexity of the situation, and makes it difficult to carry out computations between points in different zones. In such situations, it may be that using a projected coordinate system is no longer appropriate.

3. Any additional properties would usually be considered after the scale factor. It may be required, for example, that the appearance of the graticule should be as simple as possible, or that meridians should be parallel to each other. In some circumstances this might be considered more important than scale factor: for example, the Mercator projection is used in navigation because of the parallel

meridians, even though it has a higher scale factor distortion than some other projection methods.

Hints as to which projection is most likely to achieve the above goals are given in the descriptions of the projection methods in the following chapters. For example, the transverse Mercator projection is suitable for regions that are longer in their north–south extent than their east–west extent, and conic projections are suitable for mid-latitude regions with a large extent in longitude. An example of the thinking involved in this process is given in the case studies in sections 13.2 and 13.3.

8
Cylindrical projections

8.1 Cylindrical equidistant projection

Following the rules and procedures outlined in Chapter 7, a cylindrical equidistant projection is formed by bringing a cylinder into contact with the sphere and 'peeling' the meridians off the sphere and onto the cylinder without any distortion. This maintains the feature that $k_M = 1$. In so doing, it is necessary to stretch each parallel of latitude to be the same size as the equator. If the circumference of the equator, L_{eq}, is given by

$$L_{eq} = 2\pi R \tag{8.1}$$

where R is the radius of the spherical Earth, and the circumference of a parallel of latitude ϕ is given by

$$L_{\phi} = 2\pi R \cos\phi \tag{8.2}$$

then, by the original definition of scale factor in equation (7.1),

$$k_P = \frac{2\pi R}{2\pi R \cos\phi} = \frac{1}{\cos\phi} = \sec\phi \tag{8.3}$$

An example of this projection for the European region is shown in Fig. 8.1. The features to be noted are:

- As with all normal cylindrical projections, the meridians are straight and parallel to each other.
- The distances along the meridians are undistorted.
- The scale along the equator is true, but the scale of all other parallels becomes increasingly distorted towards the poles, with the extreme case of the poles themselves being represented as straight lines. Here it should be noted that

  $$\sec 90^\circ = \infty \tag{8.4}$$

  In consequence, the shape and the area become increasingly distorted towards the poles.

- The flat, square appearance of the graticule leads to the French term *plate carrée*, which is sometimes also used in English.

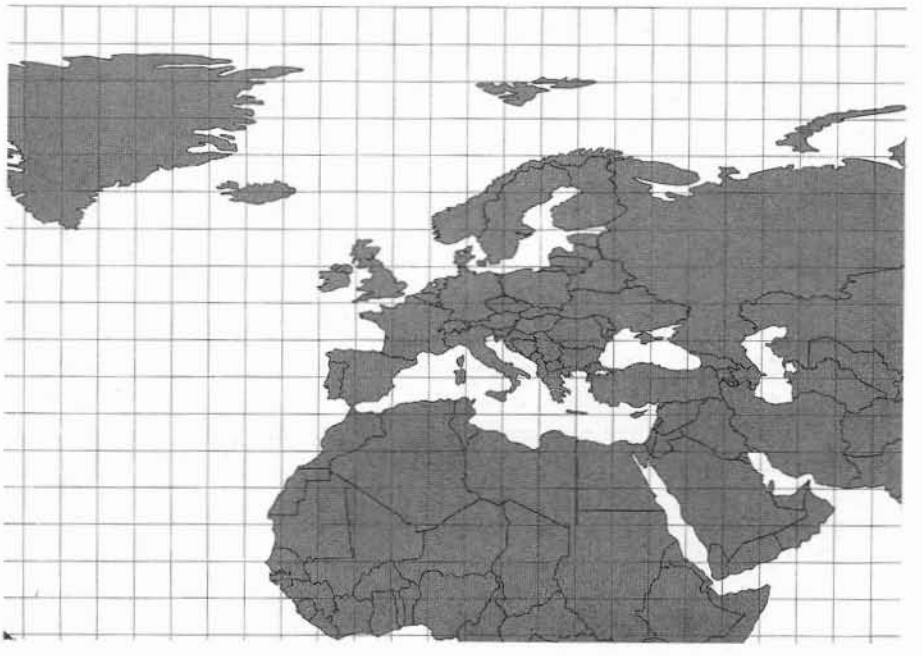

Figure 8.1 Cylindrical equidistant projection.

In computational terms, it is necessary to have a set of formulae to determine the coordinates of points on the map (E,N) given their geodetic coordinates. The first step is to select an *origin for the projection*. Here it may be conveniently chosen, for example, as

$$\phi_0 = 30^\circ \qquad \text{(the origin of latitude)}$$
$$\lambda_0 = 0^\circ \qquad \text{(the origin of longitude)}$$

The x or E (eastings) coordinate is calculated as the distance along the equator between the projected point and the origin, or

$$E = R\Delta\lambda \tag{8.5}$$

where

$$\Delta\lambda = (\lambda - \lambda_0) \tag{8.6}$$

and is expressed in radians.

Similarly, the y or N (northings) coordinate is expressed as the distance along a meridian from the origin to the projected point, or

$$N = R\Delta\phi \tag{8.7}$$

where

$$\Delta\phi = (\phi - \phi_0) \tag{8.8}$$

This then has the effect that any points south of 30° N or west of 0° will have negative coordinates on the projection. This is generally undesirable, and is avoided by adding a suitably large number to all the coordinates obtained so far. Thus:

$$E' = E + E_0 \tag{8.9}$$

$$N' = N + N_0 \tag{8.10}$$

where E' and N' now represent the *false eastings* and *false northings* of the point and E_0 and N_0 are the *false eastings* and *false northings* of the origin, which are parameters to be defined for the projection. The original coordinates E and N are referred to as the *true eastings and northings*.

It is also possible to derive the geodetic coordinates of any point whose projection coordinates are known:

$$\phi = \phi_0 + \frac{N' - N_0}{R} \tag{8.11}$$

$$\lambda = \lambda_0 + \frac{E' - E_0}{R} \tag{8.12}$$

Several important points should be noted.

- These formulae are specific to the cylindrical equidistant projection. They have been written in full as an example of the information that is necessary for defining a projection: it is not possible in a volume such as this to quote full formulae for all projections. It is to be assumed that the reader will have access to the necessary formulae via software packages such as Arc/Info. Further comments on this are given in Chapter 11.
- The formulae for a spheroid are a little more complex than those given above, which were derived using a spherical Earth.
- The concepts of the *origin of the projection* and the *false coordinates* have for convenience been introduced with reference to the cylindrical equidistant projection. It should be noted, however, that these are applicable to *any* projection, and are a part of the set of parameters that define it.

8.2 Cylindrical equal area projection

It was shown in the previous section that the cylindrical equidistant projection distorts parallels by the scale factor $\sec\phi$ while leaving the meridians undistorted. It is therefore the case that areas have also been distorted. To compensate for this, an equal area projection can be formed according to the rule of equation (7.4) that $k_M k_P = 1$. Since

the parallels must be distorted by $\sec\phi$ whatever happens (to fit onto the cylinder), this leads to the conclusion that

$$k_{\mathrm{M}} = \frac{1}{k_{\mathrm{P}}} = \cos\phi \tag{8.13}$$

Hence each small section of each meridian is multiplied by $\cos\phi$ as it is 'unpeeled' and placed on the projection. This leads to the result shown in Fig. 8.2. The features to be noted are:

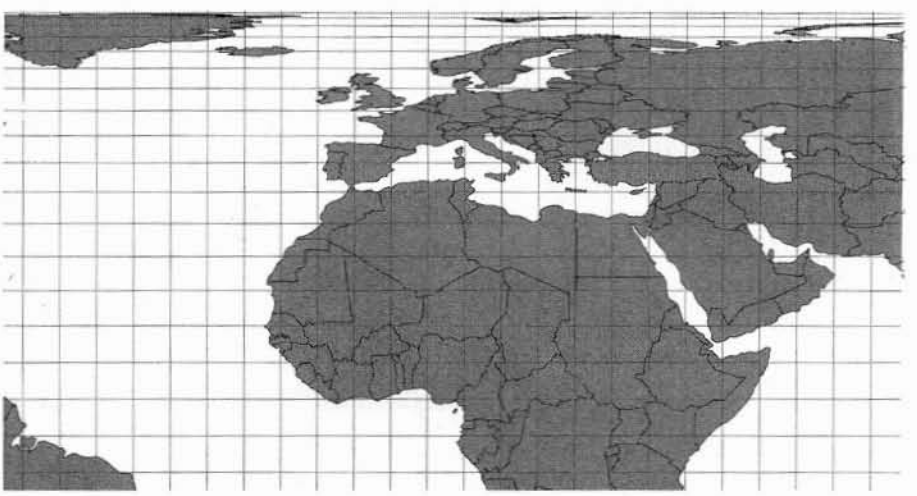

Figure 8.2 Cylindrical equal area projection.

- The scale factor in the equatorial region is close to 1 for both the meridians and the parallels. This is a consequence of cylindrical projections being optimal for the equatorial regions.
- The scale factor along the meridians is no longer equal to 1, and hence distances cannot be measured directly off such a map. Furthermore, the correction to be applied is not a straightforward one, as the scale factor is a function of latitude.
- The distortion of shape is now extreme towards the poles, as in addition to the scale factor $\sec\phi$ along the parallels there is also the distortion $\cos\phi$ along the meridians.
- Formulae can again be derived for converting between (ϕ,λ) and (E,N), which require the coordinates of the origin and the false coordinates as input.

A further refinement of this projection is to keep the equal area property but to change the shape, by applying a further scaling of 0.5 along the parallels and 2 along

the meridians. This leads to what is usually referred to as the *Peters projection*, often used by international organisations for displaying the countries of the world in their correct relative sizes. This is shown in Fig. 8.3. Note that the shape of features is now correct in the mid-latitudes, as opposed to the equatorial regions with the conventional form, and that there is less shape distortion near the poles.

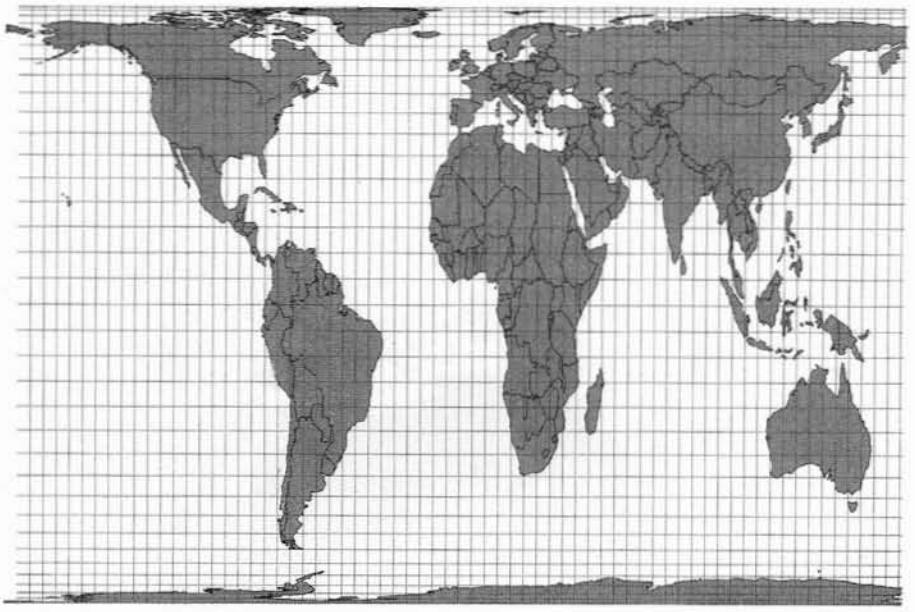

Figure 8.3 Peters projection.

8.3 Mercator projection

One of the most important of all cylindrical projections is the conformal version, which is given the particular name *Mercator*. In this projection, note again that $k_P = \sec\phi$, and hence, following equation (7.5),

$$k_M = k_P = \sec\phi \tag{8.14}$$

This then leads to a projection such as that shown in Fig. 8.4. The general features of the Mercator projection are:

- The scale factor at any point and in any direction is equal to $\sec\phi$, the secant of the latitude.
- In consequence, the pole is now of infinite size and at infinite distance from the equator, and hence cannot be represented on the projection.
- The fact that the meridians are parallel to each other and that the angles are preserved makes this an ideal projection for navigation. A line drawn between

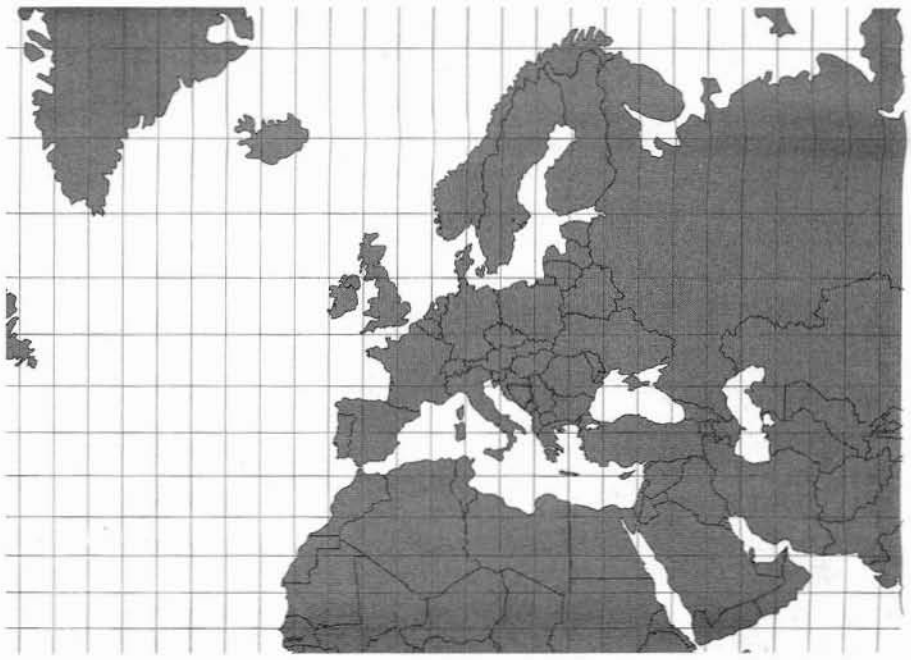

Figure 8.4 Mercator projection.

two points on the map, A and B, as shown in Fig. 8.5, has a constant angle with respect to the meridians (the *azimuth from north*), which can be read directly from the map. This is then the azimuth that should be followed in navigating from A to B. Such a line is termed a *rhumb line* or a *loxodrome*. It should be noted, however, that this line is not the shortest route between A and B, owing to the variation of scale factor within the projection. The shortest route between the two, the *great circle*, will in fact be projected in most cases as a curved line.

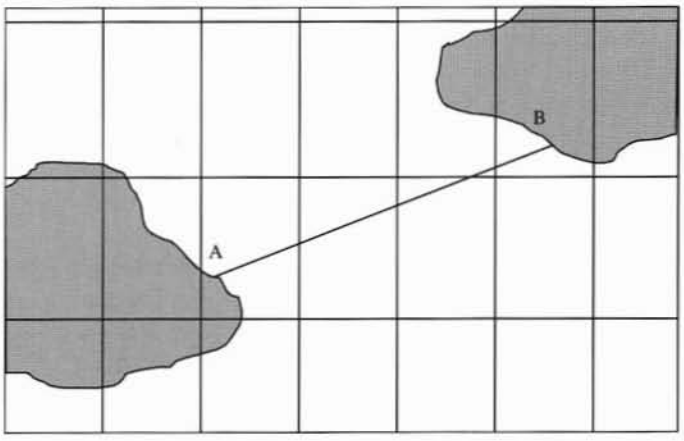

Figure 8.5 Line of constant azimuth from A to B on a Mercator projection.

- An individual map sheet (or navigation chart) usually represents a small part of the world. Hence it has a scale factor that varies within the map according to the range of latitudes it represents. A chart of the English Channel, for example, might represent the range of latitudes between 49°N and 51°N. The scale factor would then vary between $\sec 49°$ and $\sec 51°$, or 1.52 and 1.59. It is appropriate then to apply an *overall scaling* to the map so that the scale factor is on average equal or closer to 1. This will not affect the shape of the map in any way, but will make any distances read from it closer to their true values. In this example an appropriate figure might be

$$k_0 = \frac{1}{1.55} = 0.645$$

so that the scale factor is now seen to vary between $k_0 \sec 49°$ and $k_0 \sec 51°$, or 0.98 and 1.03, which means that a distance read from the map will be correct to within 3%. Because the scale factor on the equator was originally equal to 1, it is now equal to k_0. A software package such as Arc/Info, for example, achieves this by asking the user to input the latitude of the parallel at which the scale is true. If the user has a projection defined in terms of the scale factor on the equator, k_0, the appropriate parallel may be computed by knowing that the general scale factor at any point, k, is now equal to $k_0 \sec \phi$.

This concept of an overall *re-scaling* of the projection is, once again, one that has been introduced for a specific example but which has a general application to all map projections. It, or its equivalent, then forms another of the parameters required to define a projection.

8.4 Transverse Mercator projection

All of the cylindrical projections discussed so far were formed by placing a cylinder in contact with the equator and, although they are often used to portray the Earth as a whole, they are therefore optimised for use in the equatorial regions.

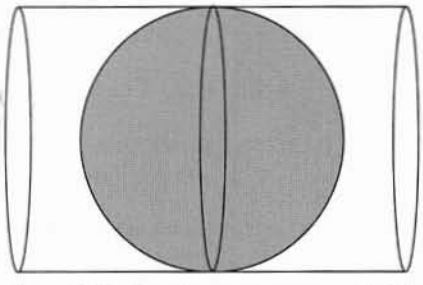

Figure 8.6 Forming a transverse cylindrical projection.

For those parts of the Earth that do not lie close to the equator, an alternative is to turn the cylinder onto its side and make the line of contact a particular *meridian*, as in Fig. 8.6. A projection so formed is termed a *transverse cylindrical* projection, and can

be based on any chosen *central meridian*. Again, a set of rules can be proposed, to produce equal area, equidistant, or conformal projections. By far the most important of these is the transverse Mercator projection, an example of which is shown in Fig. 8.7, which has been based on 0° as the central meridian. The important features of the transverse Mercator projection are:

Figure 8.7 Transverse Mercator projection centred on 0°.

- The scale factor at each point is the same in any direction, and is given by

$$k = \sec\theta \tag{8.15}$$

where θ is exactly analogous to ϕ, except that it is the angular distance from the central meridian, rather than the angular distance from the equator. It is in a sense a 'sideways version of latitude'. Note that it is not the same as longitude, but for a sphere can be found from the expression

$$\theta \approx \Delta\lambda \cos\phi \tag{8.16}$$

- The meridians are no longer parallel to each other, and in fact are no longer straight lines. The exception is the central meridian, which is a straight line. At any general point, the meridian makes an angle with the central meridian (which is also the direction of *grid north*) and this angle is termed the *convergence*, γ. For the sphere,

$$\gamma \approx \Delta\lambda \sin\phi \tag{8.17}$$

- Although Fig. 8.7 shows the transverse Mercator applied to the European region, it is more commonly used for a narrow band of no more than $\pm 3°$ on either side of the central meridian. This is principally due to its use as a survey projection, in which it is required to minimise the scale factor distortion at the expense of the extent of coverage. In this situation, the scale factor varies between 1 on the central meridian and 1.0014 on the edge of the projection, as shown in Fig. 8.8. It is then possible to minimise the scale factor distortion by once again applying an overall scaling k_0, which in the context of the transverse Mercator projection is now termed the *central meridian scale factor.* An appropriate and typical value would be 0.9996, which means that the scale factor across the projection now ranges from 0.9996 on the central meridian to 1.0010 (0.9996 sec 3°) on the edge. This is represented by the lower line in Fig. 8.8.

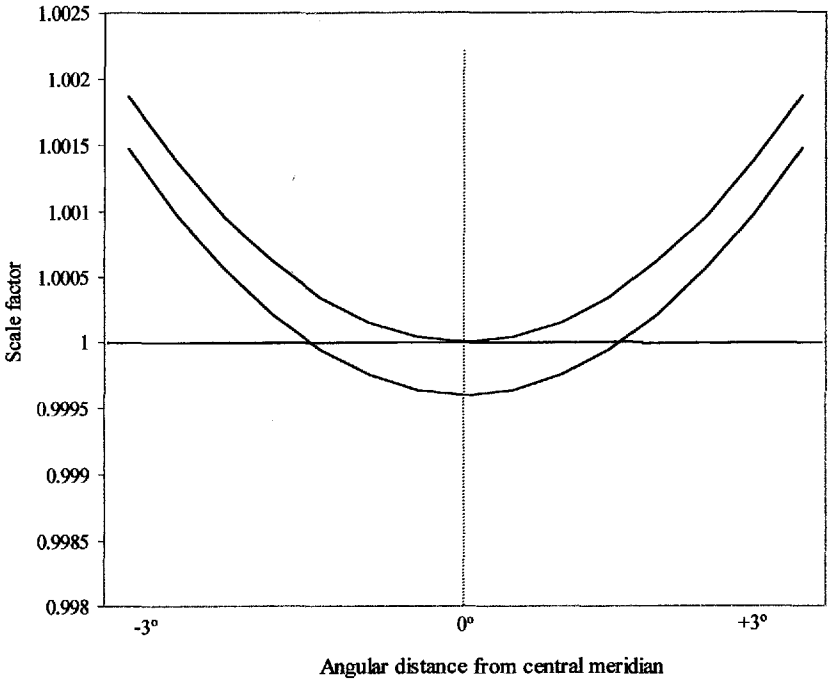

Figure 8.8 Scale factor for transverse Mercator projection.

The transverse Mercator projection is very widely used, and is particularly appropriate for regions with a large extent north–south but little extent east–west. It is, for example, the projection used for the Ordnance Survey National Grid for maps (and digital products) of Great Britain (Ordnance Survey, 1995b). In this case the relevant parameters are given in Table 8.1.

The transverse Mercator is also the basis of a worldwide projection system known as universal transverse Mercator (UTM). This system divides the world up into 60

Table 8.1 Parameters for the OS and UTM projections

Projection	λ_0	ϕ_0	False east (m)	False north (m)	k_0
OS National Grid	2° W	49° N	+400000	−100000	0.999 601 271 7
UTM ($\phi > 0°$)	Zonal	0°	+500000	0	0.999 6
UTM ($\phi < 0°$)	Zonal	0°	+500000	+10000000	0.999 6

zones of longitude, each of width 6°. The zones are numbered from 1 starting at a longitude of 180° E, and increase eastwards, as shown in Fig. 8.9. Thus the UK, for example, lies in UTM zones 30 and 31.

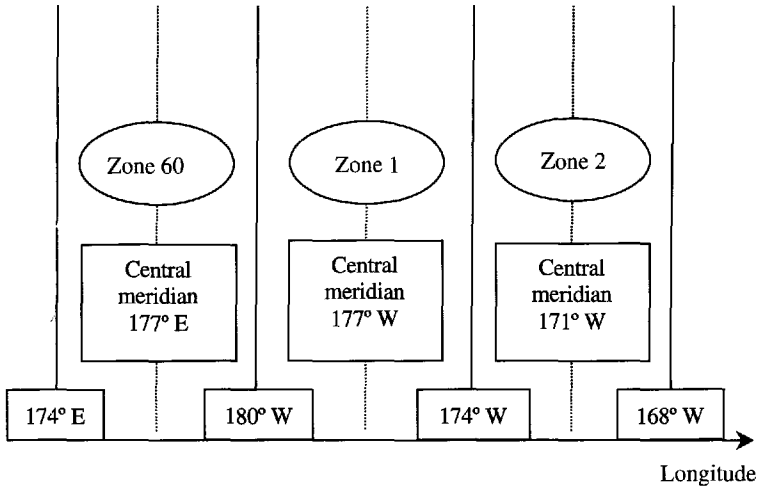

Figure 8.9 The universal transverse Mercator system.

Within each zone, the central meridian lies along the centre of the zone. Thus, UTM zone 1 has central meridian 177° W, for UTM zone 2 it is 171° W, and so on. All other parameters of the UTM system are as given in Table 8.1. Specifying that a projection is in UTM is therefore sufficient to define it completely, provided that the zone number is specified. Regions that fall across UTM zones would be shown on separate projections, with a discontinuity in between. Features falling across the boundary would have approximately the same scale factor on either side, but would be rotated with respect to each half since the convergence is in opposite directions (true north is always *towards* the central meridian).

Once again, the concept of zoning is one that has been introduced with reference to a particular projection but which is generally applicable for others. Note that the term *Gauss–Kruger* is also sometimes used for this projection.

8.5 Oblique Mercator projection

The final classification of cylindrical projections is used where a country or region to be mapped is longer in one direction than another but is not aligned along a meridian

or parallel. In this situation it is possible to formulate an oblique aspect of the Mercator projection to minimise the scale factor, as in Fig. 8.10. In defining this projection

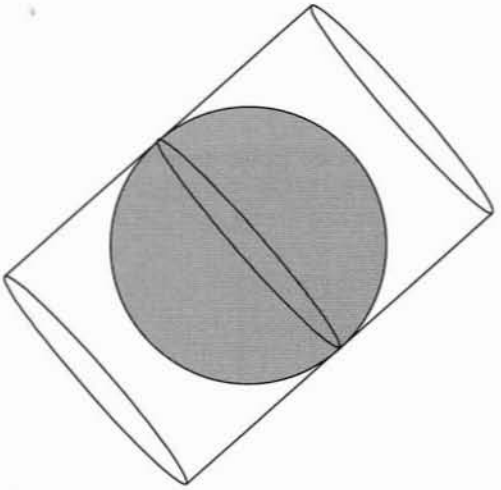

Figure 8.10 Forming an oblique Mercator projection.

it is necessary to specify the azimuth of the central line, as well as all the other parameters discussed earlier. The scale factor will now be proportional to the secant of the distance from the centre line. An example of the use of this projection is in peninsular Malaysia, where the projection is termed the Hotine oblique Mercator, or rectified skew orthomorphic.

A similar projection is the *space oblique Mercator*, which was developed for displaying satellite imagery. In this projection, the centre line is the ground track of the satellite. The formulae are complex, but again the scale factor is approximately proportional to the secant of the distance from the centre line.

9
Azimuthal projections

9.1 General azimuthal projection

An azimuthal projection is formed by bringing a plane into contact with the sphere or spheroid and formulating a set of rules for the transfer of features from one surface to the other. Once again the properties preserved can be distance, area, shape, or others.

Because the point of contact between a sphere and a plane is a single point, the scale factor distortion will be circularly symmetric. That is, the scale factor will be proportional to the distance from the centre of the projection. An azimuthal projection is therefore particularly suited to small 'circular' features on the surface of the Earth.

A special case of the azimuthal projection is where the point of contact is one of the poles. This is referred to as a *polar projection.* This has the rather obvious application of mapping the polar regions. A polar projection is formed by taking the meridians off the sphere and placing them on the plane. The amount of distortion of the meridians is a function of the type of projection, with the distortion of the parallels following in consequence. The general form of the polar projection is therefore a set of meridians radiating from the pole with no distortion of the angle at the centre. This is shown in Fig. 9.1.

As with the cylindrical projections, the azimuthal projections can have a further overall scaling applied to them, which has the effect of reducing the scale at the centre to less than 1, and making the scale true along a circle centred on the projection point. For the polar aspect this will make the scale true along a parallel of latitude away from the pole.

9.2 Azimuthal equidistant projection

The azimuthal equidistant projection is formed by keeping the scale factor equal to 1 in the direction radial from the centre of the projection. In the case of the polar equidistant projection, an example of which is shown in Fig. 9.2, this means that the scale factor on the meridians, k_M, is equal to 1. The scale factor along a parallel, k_P,

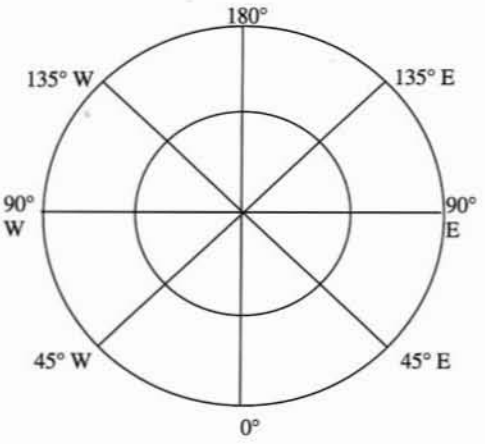

Figure 9.1 General form of the polar projection. Distances of parallels from the centre are a function of the projection type.

is given as a function of latitude, ϕ, by

$$k_P = \frac{\frac{1}{2}\pi - \phi}{\cos\phi} \tag{9.1}$$

and thus increases from 1 at the pole to 1.02 at 70° and 1.09 at 50°.

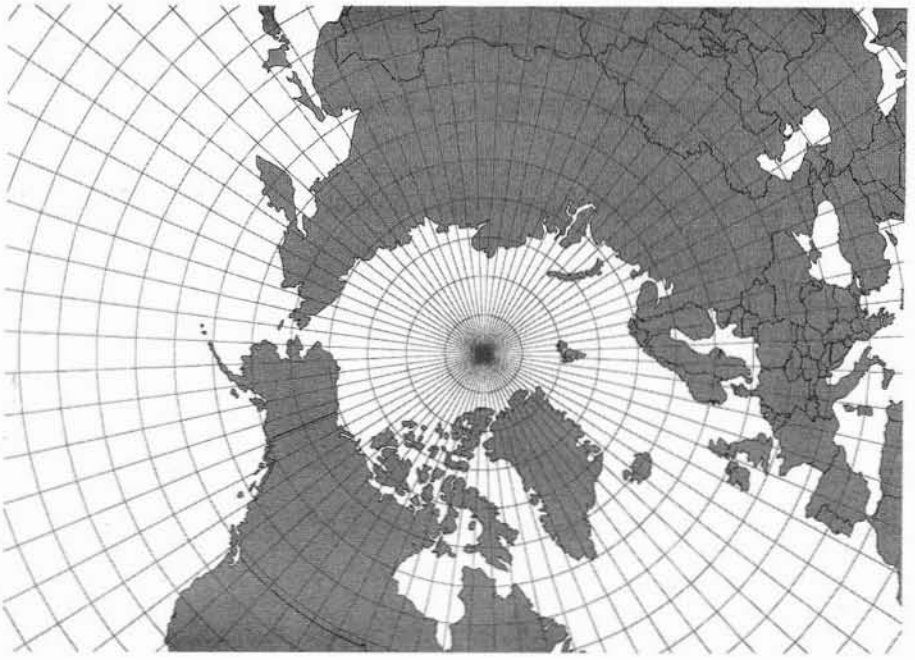

Figure 9.2 Polar equidistant projection.

As a further example, Fig. 9.3 shows an azimuthal equidistant projection that is centred on London. All distances from London are correct when measured from the map; all other distances are too long.

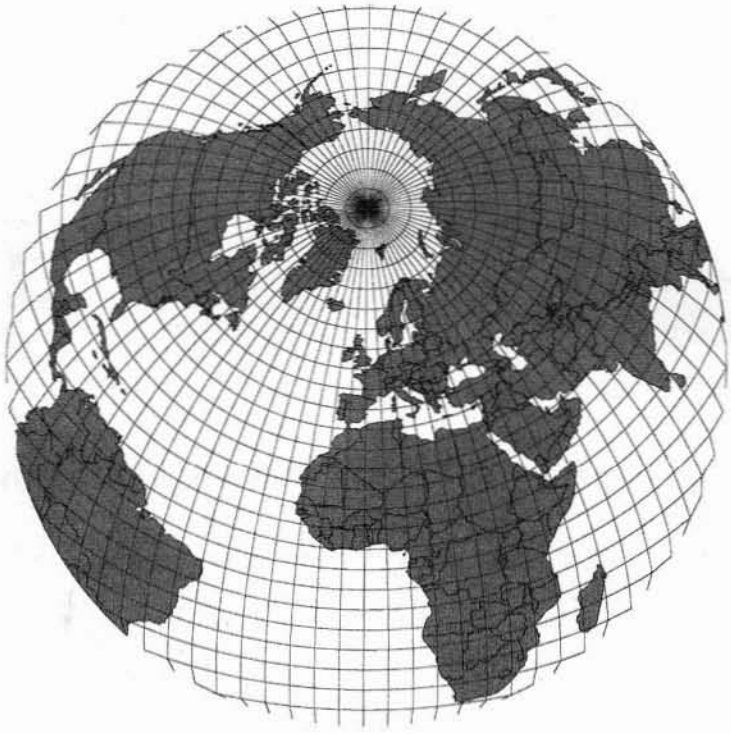

Figure 9.3 Azimuthal equidistant projection centred on London.

9.3 Azimuthal equal area projection

The azimuthal equal area projection is formed in a similar way to the azimuthal equidistant projection, except that the scale factor of the lines radial from the centre is set to the inverse of the scale factor in the perpendicular direction. For the polar aspect, shown in Fig. 9.4, this leads to scale factors of

$$k_{\mathrm{M}} = \cos(45^\circ - \tfrac{1}{2}\phi) \tag{9.2}$$

$$k_{\mathrm{P}} = \sec(45^\circ - \tfrac{1}{2}\phi) \tag{9.3}$$

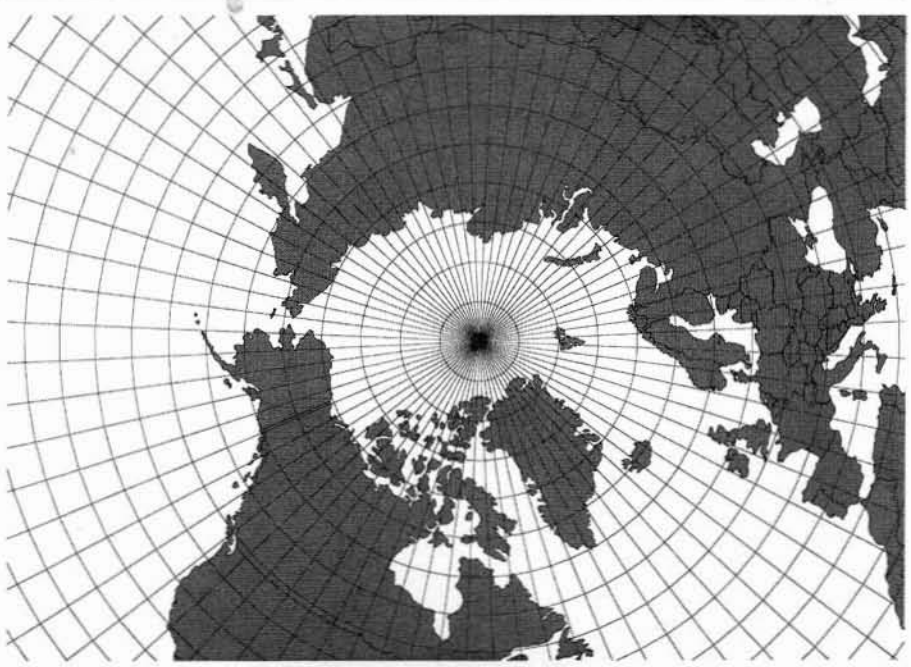

Figure 9.4 The polar equal area projection.

9.4 Azimuthal stereographic (conformal) projection

The conformal version of the azimuthal projection is termed the *azimuthal stereographic projection*. This is for historical reasons, because this projection can be constructed graphically by projecting all points from a 'viewing point' on the opposite side of the Earth from the centre of the projection.

As with all conformal projections, this one has a particular significance as it is sometimes used as the basis for national mapping, being particularly appropriate for small, compact countries or islands. The polar stereographic projection, shown in Fig. 9.5, is used as a complement to the universal transverse Mercator beyond latitudes $\pm 80°$, when it is known as the *universal polar stereographic projection* (UPS). In this usage, the scale at the pole (k_0) is reduced to 0.994, which results in a *standard parallel* (where scale is true) of $81°\,06'\,52.3''$. The false eastings and northings in this version are

$$E_0 = +2\,000\,000\,\text{m} \qquad N_0 = +2\,000\,000\text{m}$$

In general the scale factor for the polar aspect is given by

$$k = k_0 \sec^2(45° - \tfrac{1}{2}\phi) \tag{9.4}$$

which is the same in any direction.

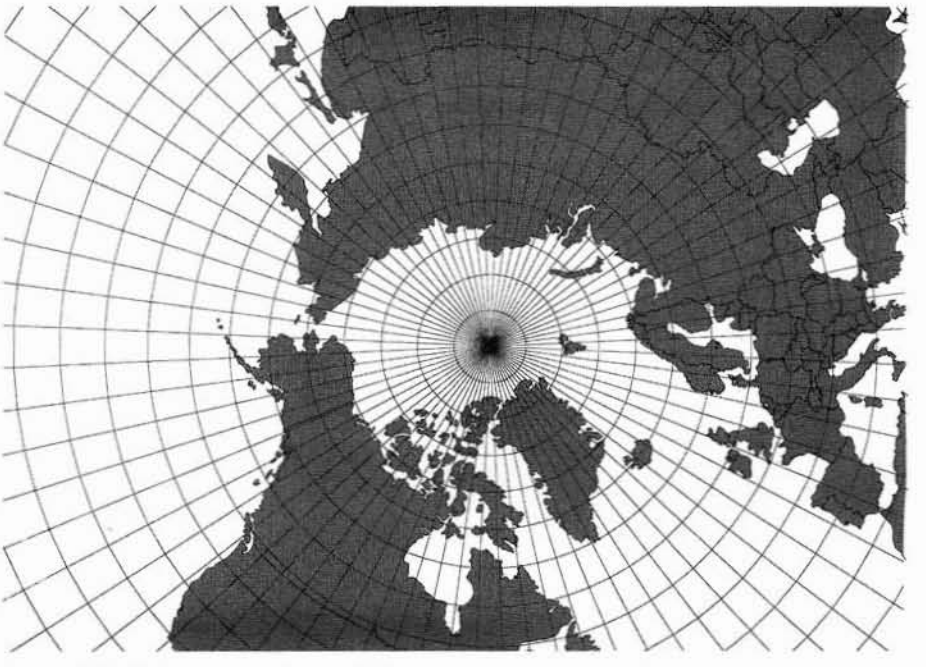

Figure 9.5 Polar stereographic projection.

9.5 Gnomonic projection

One final form of azimuthal projection should be noted here, as although it is seldom used in modern applications it does have some interesting properties. This is the gnomonic projection, which is formed by projecting all points from the centre of the Earth as shown in Fig. 9.6.

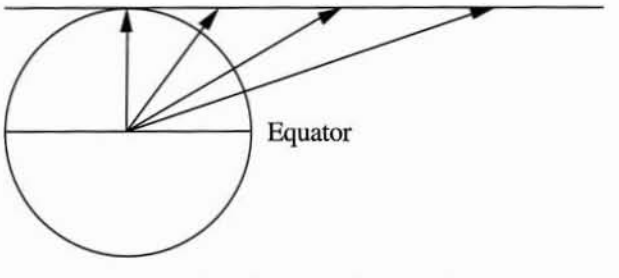

Figure 9.6 Formation of a gnomonic projection.

An example of the polar aspect is shown in Fig. 9.7. As would be expected, the scale factor distortion becomes extreme away from the centre of the projection, reaching a value of 2 along the meridian when the latitude is 45° and a value of 4 at a latitude of 30°. It is clearly not possible to show an entire hemisphere with this projection.

The only advantage of this projection is that it is the only one where all great circles (the shortest route between two points on a sphere) are shown as straight lines

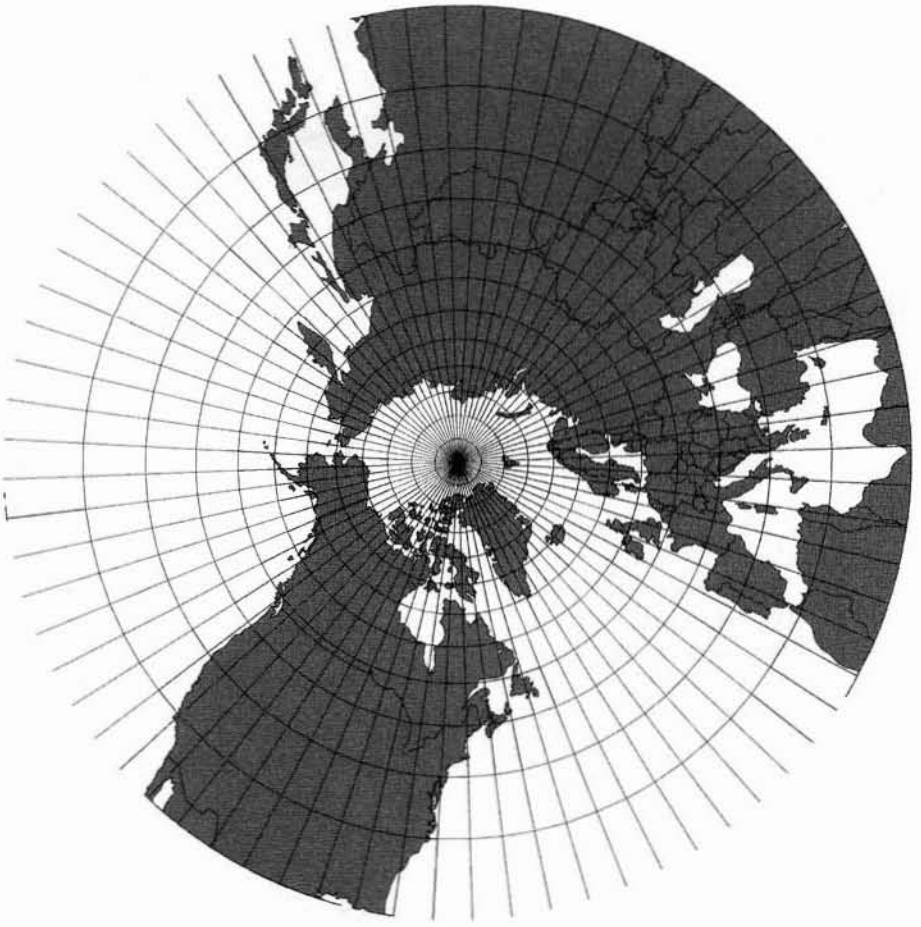

Figure 9.7 The (north polar) gnomonic projection.

on the projection, and vice versa. This feature means that it can be used to plan the shortest route between two points, although this role has largely been superseded by computational techniques.

10
Conic projections

10.1 General conic projection

A conic projection is formed by bringing a cone into contact with the sphere or the spheroid. In so doing it is seen to be touching the sphere along a parallel of latitude. This line is known as the *standard parallel* of the projection.

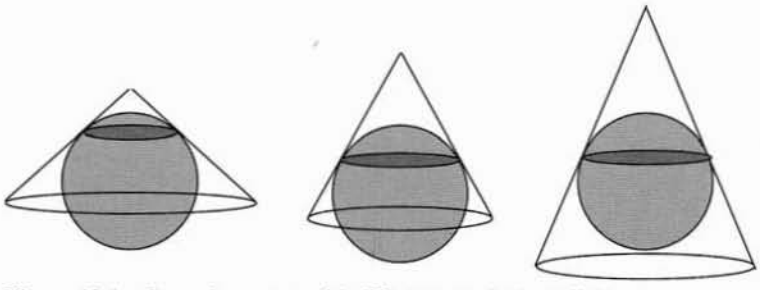

Figure 10.1 Cones in contact with different standard parallels.

It can be seen from Fig. 10.1 that many different shapes of cone can be selected, all resulting in a different standard parallel. The choice will depend upon which region of the Earth is to be mapped, an appropriate standard parallel being one that passes through the centre of the region.

The resultant form of the conic projection is that the meridians appear as straight lines converging towards one of the poles. The angle between two meridians is a function of the standard parallel, and can be expressed as

$$\gamma = \Delta\lambda \sin\alpha \tag{10.1}$$

where $\Delta\lambda$ is the difference in longitude of the two meridians and α is the latitude of the standard parallel.

The conic is in fact a general case of projection, of which the cylindrical and azimuthal projections are particular forms. In equation (10.1), as α tends towards 90°, γ tends to $\Delta\lambda$, which indicates that the angles between the meridians are true. This is as in the polar projection, which is the equivalent of a completely flat cone touching the sphere at the pole. Similarly, as α tends to 0°, γ tends to 0°, which indicates that

the meridians are parallel as in a normal conic projection. A cylinder is the equivalent of a cone touching the equator.

These considerations are useful for gaining an insight into the nature of conic projections, but should not be implemented in practice, as the formulae for the cone are likely to break down under these extreme conditions.

The equivalent of an overall scaling is often used for conic projections, where it is achieved by using *two standard parallels* as shown in Fig. 10.2. The effect is to reduce the scale factor below 1 between the two standard parallels and increase it above 1 outside them.

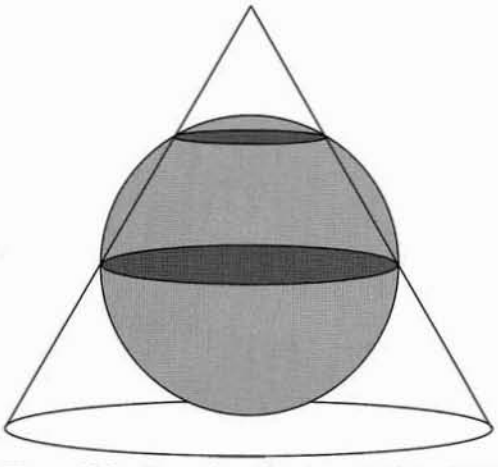

Figure 10.2 Formation of conic projection with two standard parallels.

Finally, it should be noted that for any conic projection the scale factor is entirely a function of latitude, and these projections are therefore suitable for depicting regions with a broad extent in longitude, particularly mid-latitude regions.

10.2 Conic equidistant projection

A conic equidistant projection preserves the scale factor along a meridian ($k_M = 1$). The parallels are then equally spaced arcs of concentric circles. The scale factor along a parallel of latitude is given as a function of latitude ϕ (Snyder, 1987) by

$$k_P = \frac{(G - \phi n)}{\cos\phi} \tag{10.2}$$

where

$$G = \frac{\cos\phi_1}{n} + \phi_1 \tag{10.3}$$

$$n = \frac{\cos\phi_1 - \cos\phi_2}{\phi_2 - \phi_1} \tag{10.4}$$

and ϕ_1 and ϕ_2 are the two standard parallels. If only one standard parallel is used, n in equation (10.4) is simply equal to $\sin\phi_1$.

An example of this projection is shown in Fig. 10.3.

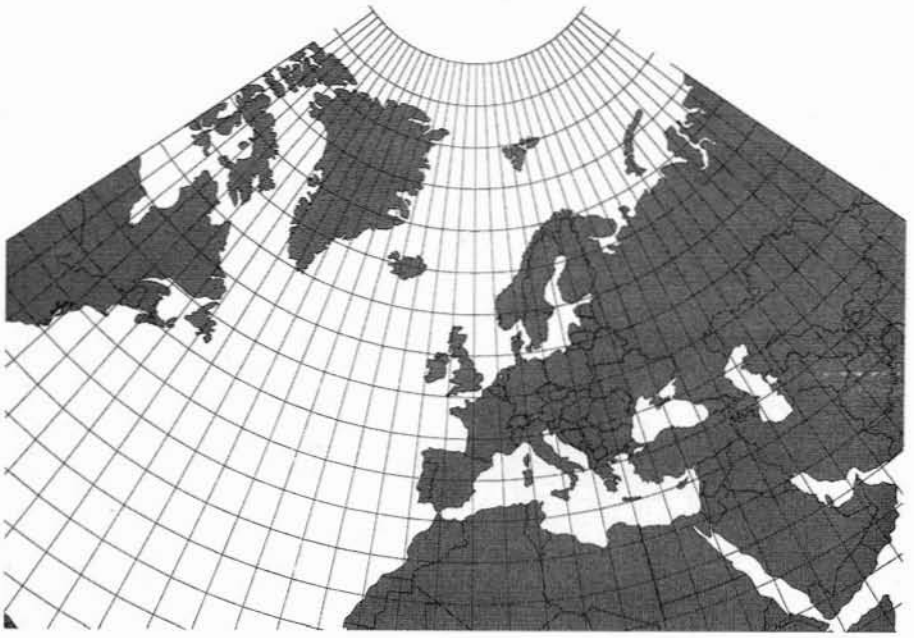

Figure 10.3 Conic equidistant projection with standard parallels at 20° and 60° N.

10.3 Albers (conic) equal area projection

The equal area version of a conic projection is usually called Albers equal area. An example of this for the European region is shown in Fig. 10.4.

It will be noted that the pole is shown in this projection as a circular arc, indicating once again that shape has been sacrificed to keep the area undistorted. It should also be noted, however, that the shape is not as badly distorted as in the cylindrical equal area projection in Fig. 8.2. This is mainly a function of the region being projected: the European area shown on these two maps is largely an area in mid-latitude with a large east–west extent, which is more suited to a conic projection than to a cylindrical one.

10.4 Lambert conformal conic projection

The conformal version of the conic projections is usually named after Lambert, who first developed it in 1772 (Snyder, 1987). The full name is the Lambert conformal conic (LCC), but most references to a Lambert projection would usually be understood to refer to this one (though not without some ambiguity). This is an extremely widely

Figure 10.4 Albers equal area (conic) with standard parallels at 20° and 60° N.

used projection, and it is probably true to say that LCC and the transverse Mercator between them account for 90% of base map projections worldwide.

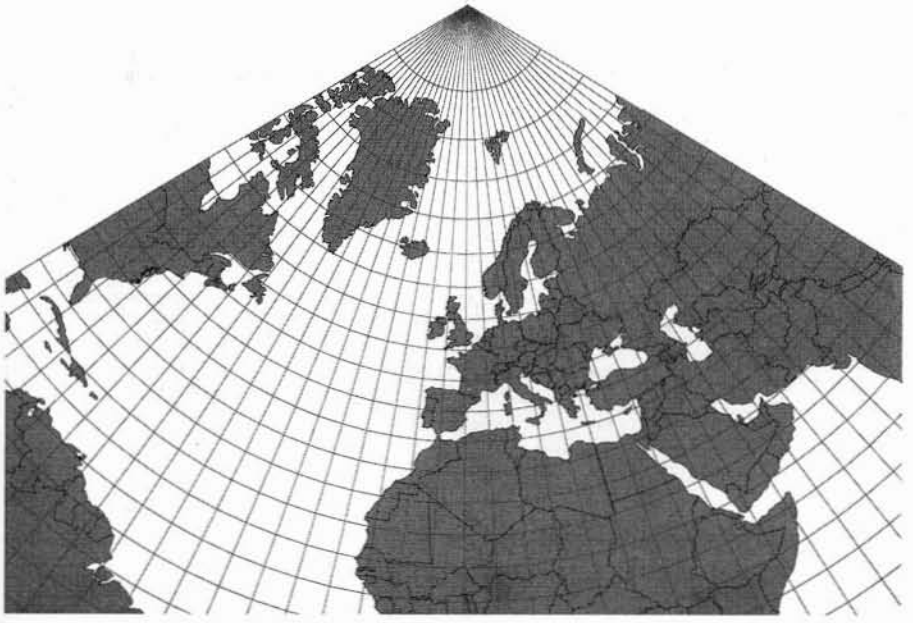

Figure 10.5 Lambert conformal conic projection with standard parallels at 20° and 60° N.

An example of the LCC is shown in Fig. 10.5. Because it is a conformal projection, the meridians meet at a point that represents the pole. The example in Fig. 10.5 was formed with the equivalent of a standard parallel at 40°. An LCC with a standard parallel at the equator would effectively be the same as a Mercator projection, with the meridians parallel and never reaching the infinite pole; one with a standard parallel at 90° would be the same as a polar stereographic projection. Examples near these extremes are shown in Fig. 10.6.

An LCC projection may also be formed with two standard parallels, as with all conic projections. In this case it is the equivalent of one standard parallel halfway between, with an additional scaling applied. The usual arrangement for minimising distortion is to have two standard parallels which are each $\frac{1}{6}$ of the range of latitude in from the extremes of the projection.

The formulae for LCC are complicated, but the expression for scale factor for the case with one standard parallel and a spherical Earth can be quoted (Snyder, 1987) as

$$k = \frac{\cos\phi_1 \tan^n\left(\frac{1}{4}\pi + \frac{1}{2}\phi_1\right)}{\cos\phi_1 \tan^n\left(\frac{1}{4}\pi + \frac{1}{2}\phi\right)} \tag{10.5}$$

where ϕ_1 is the standard parallel and $n = \sin\phi_1$.

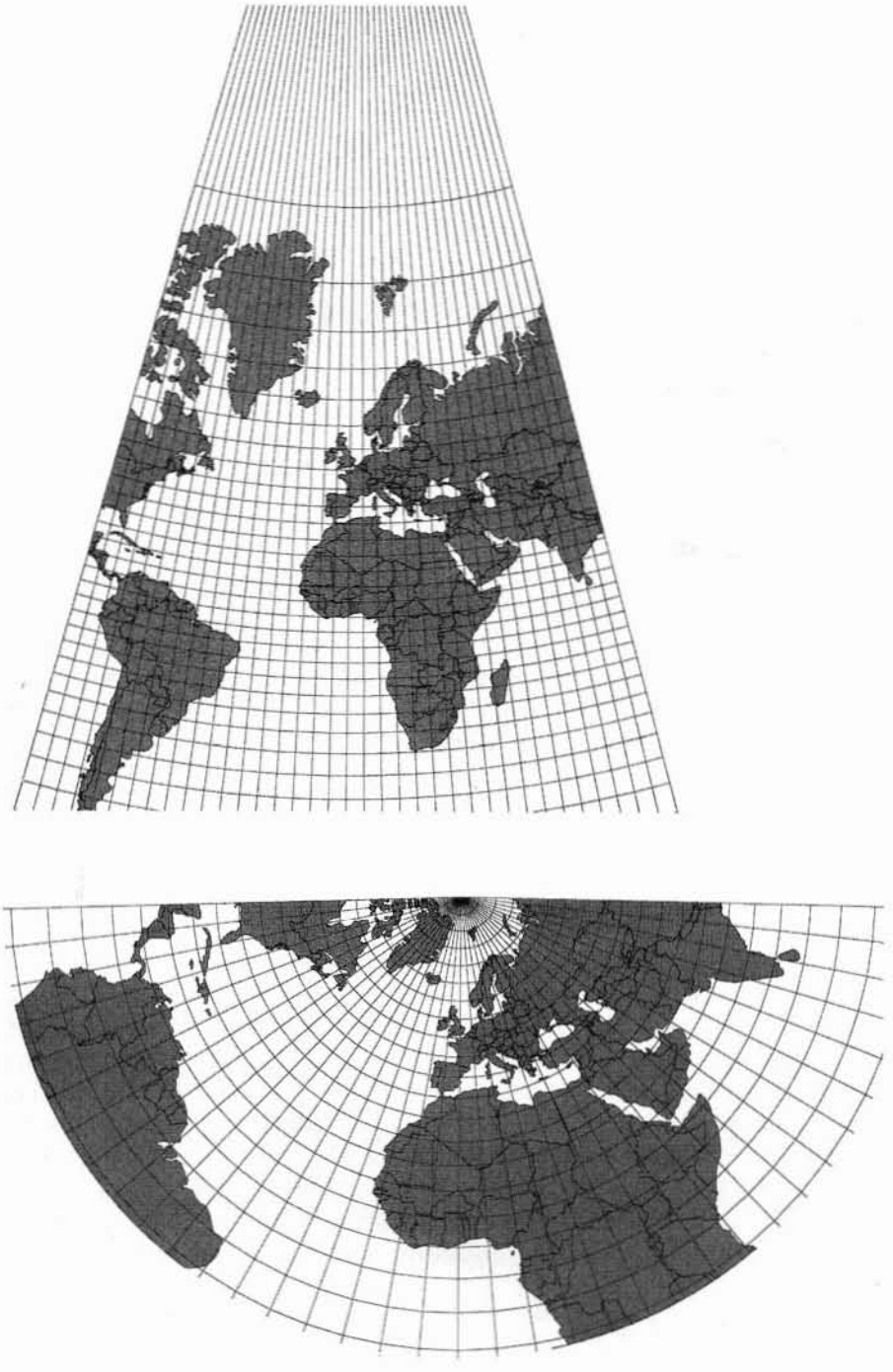

Figure 10.6 LCC projections at 80° N (above) and 10° N (below).

11
Summary of information required

11.1 Formulae

Examples of formulae for converting between geodetic coordinates and a projected coordinate system were given in section 8.1 for the cylindrical equidistant projection. Throughout this book, however, the assumption has been made that the vast majority of users will not be called on to program specific algorithms into a computer, but will have access to a wide variety of projections via a suite of software. ArcView, for example, supports over 40 different projections. Some of these – for example UTM, UPS, and the British National Grid – are true projected coordinate systems and include the parameters in the specification; others are generic methods of projection and require parameters as input. In such cases, the user will be restricted to choices of sub-sets of these projections, being those defined by the particular parameters summarised in section 11.2.

It may occasionally prove useful to have access to a set of programs for projection computations that are independent of a particular application. One such is the PROJ.4 package from the United States Geological Service, which is available over the World Wide Web (USGS, 1999).

If it is necessary to use a particular projection for which no computer program is available, a standard work of reference must be consulted. One of the best available is Snyder (1987), which has already been referred to in some of the formulae given in preceding sections. This work contains formulae in spherical and spheroidal form for over 30 main types of projection, and includes a comprehensive set of worked examples for each.

11.2 Parameters

Each projection is defined in terms both of the formulae needed for converting between geodetic and grid coordinates, and of the parameters of the projection which are necessary as input for those formulae.

Except for standard projections such as UTM and UPS, a projection is not defined simply by its type. A transverse Mercator projection with central meridian 2° W, for example, is completely different from a transverse Mercator projection with central meridian 20° E.

Does it matter what the parameters are and what projection has been used? A map may be digitised in projection coordinates and used in much the same way that a paper map would be used. However, as soon as it is necessary to carry out any kind of computation based on the information presented in the map, a knowledge of the projection and its parameters is necessary. This may be a simple computation such as the area of a feature on the map or the distance between two points, in which case the projection may be ignored for low accuracy applications. But if it is required to combine the data from the map with other data, perhaps from a satellite image or from additional information obtained and referenced using GPS, it is necessary to transform all information into a common coordinate system. For this, a knowledge of the projection and its parameters is required.

Table 11.1 summarises the information that is usually necessary for each category of projection. The meaning of these terms is summarised in Table 11.2.

Table 11.1 Parameters necessary to define each projection

Projection	ϕ_0	λ_0	E_0	N_0	k_0	A_0	ϕ_1	ϕ_2
Cylindrical equidistant	✓	✓	✓	✓	✓			
Cylindrical equal area	✓	✓	✓	✓	✓			
Mercator	✓	✓	✓	✓	✓			
Transverse Mercator	✓	✓	✓	✓	✓			
Oblique Mercator	✓	✓	✓	✓	✓	✓		
Azimuthal equidistant	✓	✓	✓	✓	✓			
Azimuthal equal area	✓	✓	✓	✓	✓			
Azimuthal stereographic	✓	✓	✓	✓	✓			
Gnomonic	✓	✓	✓	✓	✓			
Conic equidistant	✓	✓	✓	✓			✓	✓
Albers equal area	✓	✓	✓	✓			✓	✓
Lambert conformal conic	✓	✓	✓	✓			✓	✓

Table 11.2 Meanings of terms in Table 11.1

ϕ_0	Latitude of the origin. Not necessarily the same as ϕ_1 for the conic projections
λ_0	Longitude of the origin. Equivalent to the central meridian for cylindrical and other projections
E_0	False eastings to be added to all coordinates. Equivalent to the eastings at the origin. May alternatively be referred to as x_0
N_0	False northings to be added to all coordinates. Equivalent to the northings at the origin. May alternatively be referred to as y_0
k_0	Overall scaling factor to be applied. May be referred to as the central meridian scale factor for transverse Mercator. Not usually applied to conic projections, as its role is performed by using two standard parallels
A_0	The azimuth of the centre line for oblique projections
ϕ_1	The first (or only) standard parallel
ϕ_2	The second standard parallel

The question then remains of where to obtain this information. Some maps have the name of the projection written in the margin, either specifically or in rather a vague way (e.g. 'conic with two standard parallels'). It is very unusual, however, to see the actual values of the parameters printed on a map.

Most national mapping organisations publish the values of the parameters used for their own map series, for example Ordnance Survey (1995b), but obtaining the information may be a lengthy process. Nor does it mean that all maps of that country will be printed in that projection: a publisher may have devised their own, and it is also difficult to obtain this information. Again, a reference such as Snyder (1987) may prove useful, though it is by no means comprehensive on this point. An alternative is Jones (1990).

If it is not possible to obtain any information on the projection used, recourse may be had to the techniques outlined in Chapter 12.

12
Direct transformations

12.1 Compatibility of coordinate systems

A frequently used alternative when two sets of data have to be combined into a single datum and projection is to transform one into the other using points that are common to both sets. If, for example, a satellite image is to be transformed into a ground system which is represented by a given map, this may well be the case. In these circumstances the projection and datum of the map are being adopted as the reference system, without having any detailed information on what its parameters actually are.

The procedure in the case of transforming an image to a map is often referred to by remote sensors as *warping*. In fact, this is an approach that is applicable in other areas as well, including the case where high precision GPS data is warped to a local grid over a small area (as discussed in section 6.5) and in GIS.

To begin with, however, it must first be considered whether these techniques are likely to produce a satisfactory result. That is, the limitations of such an approach will be explored before dealing with the detailed procedures. It is easy to demonstrate situations where any kind of two-dimensional transformation is not possible, by reference to two absurdly different projections (see Fig. 12.1). We are therefore concerned

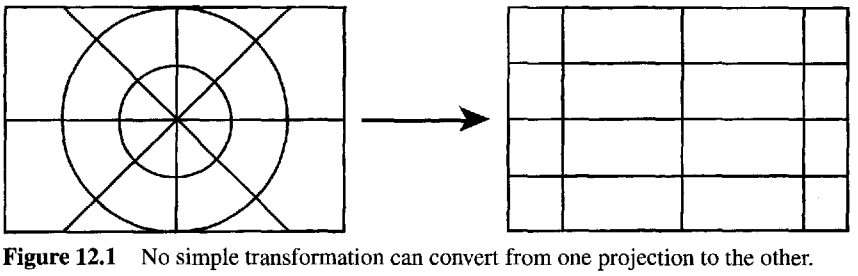

Figure 12.1 No simple transformation can convert from one projection to the other.

with a consideration of how close the *shapes* of the two projections are. For some situations, as in Fig. 12.1, the answer is obvious. It might be thought that it is always possible to transform from one conformal projection to another, since by definition the shape is preserved. This condition only holds true for small features, however, because the scale factor varies within the projection and it is therefore impossible for the shape of a large area to be preserved. An example of this is shown in Fig. 12.2,

which represents the effect of a variation in scale factor within a projection on a square of large dimensions.

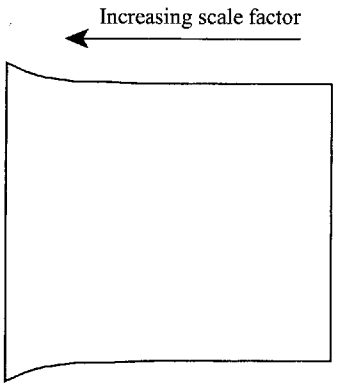

Figure 12.2 Scale factor variations within a projection.

The point of concern, then, is the variation in the ratio of scale factor between the two projections across the area concerned. If this figure is largely constant, a simple overall scale applied to one image or projection is likely to account for most of the differences between the two. If, on the other hand, there is a large variation across the area, it is questionable whether a simple transformation is likely to be adequate. The key parameter to be considered here is the ratio of the scale factors between the two projections and its variations within the area under consideration. This parameter can be defined as

$$r = \frac{k_A}{k_B} \tag{12.1}$$

where k_A is the scale factor in the target projection and k_B is the scale factor in the source projection, and r therefore represents the scale change in going from one projection to the other. If r is more or less constant across the area concerned, the two projections have essentially the same geometry, and a simple two-dimensional similarity transformation is likely to yield satisfactory results.

It is of course necessary to be a little more precise than the statement 'more or less constant', however, so consider a situation in which the scale factor ratio across an area ranges from a minimum of r_{min} to a maximum of r_{max}. A similarity transformation applied to the data set as a whole would in effect apply a mean scale factor ratio of r_{mean} (that is, assuming that the control points used to derive the transformation were evenly spread out over the whole area). The situation might, for example, look like the one depicted in Fig. 12.3, in which the scale factor ratio increases from a minimum down the left-hand side to a maximum down the right-hand side. If an overall scaling is applied in this example, the right-hand side of the area will be scaled along with all other data by the ratio r_{mean} instead of the correct ratio r_{max}. The mismatch along the

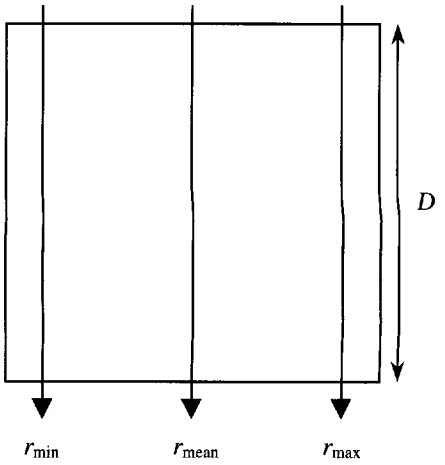

Figure 12.3 Variation of scale factor ratio.

right-hand side of the image will then be given by

$$\varepsilon = D\left(\frac{r_{max}}{r_{mean}} - 1\right) \tag{12.2}$$

where D is in this case the length of the right-hand side, or more generally the dimension of the region where the incorrect scale factor is being applied. Equation (12.2) therefore provides a reasonably good rule of thumb for the errors, ε, that are likely to result from the application of a similarity transformation.

This concept can be extended to a consideration of higher order transformations, as covered in sections 12.2–12.4. These can be thought of in conceptual terms as 'tying' two data sets together at a set of control points that are identifiable in both systems. Exactly the same formula may be applied, but in this situation the extremes r_{min} and r_{max} are to be thought of not across the whole area but between two control points. The dimension D is then the distance between the control points, rather than the dimension of the whole area. This is illustrated in Fig. 12.4.

Equation (12.2) will under these circumstances provide a rather conservative estimate of the likely error, as it takes no account of the non-linearity of the transformation. However, it does assume that a transformation appropriate to the number of control points is being used. That is, there is nothing wrong with using half a dozen control points to derive the parameters of a similarity transformation (in fact it is laudable), but it should not be assumed that the projection problems have been overcome as a result.

A numerical example of an assessment of the compatibility of two data sets is given in the case studies in sections 13.4 and 13.5.

12.2 Ground control

In order to determine the parameters of plane transformations (section 12.3) or the unknowns in equations (12.3)–(12.13), it is necessary to use ground control points

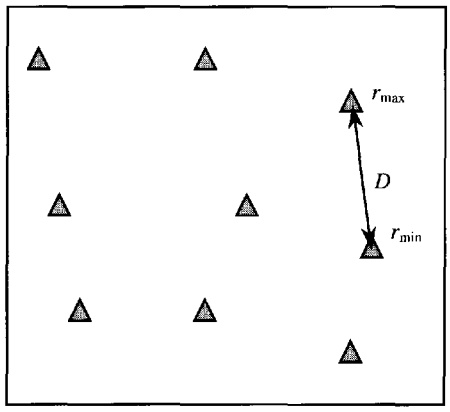

Figure 12.4 Control for higher order transformation.

(GCPs). A GCP must be recognisable in both data sets and must have known coordinates in a suitable ground reference system. The number of points required depends on the method used and is discussed in the relevant sections. GCP coordinates may be obtained directly by survey measurement or from a map. In either case the coordinates will be given in a reference system; this may be geodetic (latitude and longitude) or cartesian (X, Y, Z) and it may be global based on the centre of the Earth, or local based on a national or regional projection. It is always important that all coordinates are given in the same system. Direct survey measurements may come from a survey based on a local or national coordinate system, which in Great Britain will be the Ordnance Survey National Grid, or they may come from the GPS.

Maps should be used with caution for determining GCPs. Map data at scales of 1 : 25 000 and smaller is notoriously unreliable because of the many errors that may have accumulated in the map production and map digitising process. These include:

- *survey errors* (in some parts of the world published maps may be based on topographical sketches)
- *drafting errors*
- *generalisation*
- *paper distortion*
- *errors in digitising* a paper document for use in the validation process.

It is always necessary to take the accuracy of the data into account when using ground control; it is particularly important when using map data.

12.3 Plane transformations

12.3.1 Applications and terminology

Distortions may be present in satellite images. If the relief of the ground is high or oblique images are being used, techniques for correction must take into account the

relief of the ground. For areas of low relief or for low resolution sensors, simpler methods may be used. Similarly, any two digitised maps may be simply transformed onto one another if the internal scale factor distortions are low. This section deals with the correction of two-dimensional data.

The correction of data in two dimensions may be approached by applying a transformation to the data and resampling it to produce a corrected image which gives a best fit to the ground control used. The transformation may be based on a theoretical consideration of the errors involved or selected on empirical grounds. The latter is the method most commonly used to produce an image which is corrected to fit to a given map projection.

A number of transformations are widely used and a brief description of the common ones is given here. The terminology used here will be that (x, y) is the coordinate system before transformation, and (X, Y) is the coordinate system required after transformation. For remote sensing applications, (x, y) represents an image coordinate system; alternatively, it might represent GPS data in WGS84 that is to be transformed into a local system. The coordinates (X, Y) are referred to as a *ground coordinate* or *local system.*

12.3.2 Two-dimensional similarity transformation (four parameters)

This transformation is used to relate any two-dimensional rectangular coordinate system to any other two-dimensional rectangular coordinate system. It preserves the internal geometry of the transformed system, so it is ideal for comparing the geometry of any two systems simply by determining the residuals and the root mean square errors after transformation. For a given control point, this transformation is defined by two equations:

$$X = ax - by + c \tag{12.3}$$

$$Y = bx + ay + d \tag{12.4}$$

Effectively, this transformation is performed by applying a scale factor, m, where

$$m = \sqrt{a^2 + b^2} \tag{12.5}$$

a rotation angle, α, where

$$\tan\alpha = \frac{b}{a} \tag{12.6}$$

and two translations (c and d).

The operations applied by this transformation (scale, rotation, and translations) are the equivalent of projecting an image by any ordinary photographic enlarger onto a map. It is used to give initial coordinates to the centre of a frame.

Similarly, the parameters of the transformation may be derived in a geographic information system such as Arc/Info if common points can be identified between two data sets. The parameters so derived are then applied to all the other points in the data set to be transformed.

12.3.3 Two-dimensional affine transformation (six parameters)

The mathematical relationship for an affine transformation may be expressed by the following equations:

$$X = a_0 + a_1 x + a_2 y \tag{12.7}$$

$$Y = b_0 + b_1 x + b_2 y \tag{12.8}$$

An affine transformation enables an adjustment to be applied independently in each direction, and is thus able to correct effects that have actual physical causes. Thus, for remotely sensed scanner images, it corrects first-order distortions such as affinity due to non-orthogonality and scale difference between scans and along track directions which may be caused by Earth rotation and other geometric distortions. For digitised maps it is able to correct for effects such as map shrinkage of a different size in each direction.

At the same time, this kind of transformation may be able to correct for some of the error caused not by physical effects but by differences of datum and map projection between an image and a map or two different maps. It should be noted, however, that it is not possible to isolate the magnitudes of physical and non-physical causes unless calculations such as that outlined in section 12.4 have been carried out.

This transformation applies scale factors in the x direction (or scan direction for a satellite image) of

$$m_x = \sqrt{a_1^2 + b_1^2} \tag{12.9}$$

and in the y direction (or flight direction for an image) of

$$m_y = \sqrt{a_2^2 + b_2^2} \tag{12.10}$$

as well as a factor of affinity:

$$F_a = \frac{mx}{my} \tag{12.11}$$

These may be determined by using ground control points or points common to both data sets. Three ground control points, at least, are required.

12.3.4 Second-order polynomials (twelve parameters)

Polynomials in the form:

$$X = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 y^2 + a_5 xy \tag{12.12}$$

$$Y = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 y^2 + b_5 xy \tag{12.13}$$

are used for correction of scanner data. If polynomials are used, great care must be taken to ensure that a sufficient number of control points are available and that they are distributed over the whole area to be transformed, as these transformations can behave in an extremely unstable manner.

A minimum of six GCPs is required to determine the transformation parameters, although it is desirable to have more to build in checks. In addition to first-order distortions, polynomials correct second-order distortions in satellite images caused by pitch and roll, sub-satellite track curvature and scan line convergence due to Earth rotation and map projection. They may also correct some of the distortions related to the attitude variations along the flight path. Additional terms may be added to equations (12.12) and (12.13) to correct for higher-order distortions; the need for care in the use of control points is greater for higher orders.

This type of transformation is also used in geographic information systems such as Arc/Info and ArcView to relate data sets that do not match after simple or affine transformations have been carried out. The use of the equations is of course invisible to the user; it is simply necessary to identify the minimum number of common points. Again, care must be taken to use points that are well distributed over the area concerned. Problems can arise if two data sets of very different geometry are being linked in this manner, as the equations can become unstable. In such situations, recourse may be had initially to the techniques outlined in section 12.4 as a first step.

This section has largely dealt with the use of two-dimensional transformations in the *forward* sense: that is, equations have been quoted that will permit the transformation to be carried out if the parameters are known. The use of the *reverse* sense has been implied: that is, the parameters must first be determined by least squares from a knowledge of the coordinates in both systems of a set of common points. The mathematical treatment of this subject is dealt with in Appendix section A3.

12.4 Unknown projections: measuring from maps

If the parameters of a projection are known, and the projection type is fully supported by the software package being used, there is not going to be a problem in transforming between projections. Likewise, it has been shown in the preceding section that, even if the parameters of the projection are not known, a simple transformation may be carried out if there is no significant change of shape between the 'source' and the 'target' data sets.

There remain some situations, however, in which the two data sets are not sufficiently similar in shape, and the parameters of the projection are not known. In such cases it may be possible to determine the parameters of the projection by an inspection of the map itself. In the argument that follows it will generally be assumed that the sphere, rather than the spheroid, is an adequate model for representing the coordinate system. If the spheroid was actually used in determining the projection coordinates this may lead to errors, but in most cases they will be errors that can be dealt with through a simple affine transformation using common points.

The approach of this section is to 'unravel' the most serious distortions caused by the projection, and to bring the coordinate system at least approximately into line with the geometry of the 'target' data set. The first requirement is that the graticule should be visible in the map, either explicitly printed or identifiable by tick marks around the edge and in the middle. Without the graticule the map is essentially an abstract concept, existing purely in projection space. On most large scale maps the grid will

be printed. If the grid is not shown, this is not critical; a user could create a grid at an appropriate scale.

The first step in treating a problem of this kind is to identify the type of projection by examining the graticule. Clues may be obtained by reading the sections on individual projection types, but some of the main features are listed below.

- *Parallel meridians aligned with grid north:* these indicate that the map is in a cylindrical projection. Within this category, further inspection may reveal the following sub-categories:
  - *evenly spaced parallels of latitude:* a cylindrical equidistant projection, as there is no change in scale along the meridian
  - *parallel spacing decreases towards the poles:* an equal area projection (as the meridional scale factor is decreasing to take account of the increased scale factors along parallels of latitude as the poles are approached)
  - *parallel spacing increases towards the poles:* the Mercator projection (by the converse argument to the one above, indicating conformality).
- *Straight meridians converging, though not at a true angle; circular parallels:* that is to say, that the angle between any two meridians is not equal to (in fact is less than) the angular difference between their longitudes. This indicates that the projection is conic. Within this category, the following sub-sections apply:
  - *evenly spaced parallels:* a conic equidistant projection (again, a constant scale factor along the meridians indicates an equidistant projection)
  - *parallel spacing decreases away from the standard parallel:* a conic equal area projection. (If there are two standard parallels, the relevant fact is whether the spacing decreases away from the mean of the two. See below for an indication of how to determine the value of the standard parallel without prior knowledge.)
  - *parallel spacing increases away from the (mean of the) standard parallels:* the Lambert conformal conic projection.
- *Straight meridians converging at a true angle between meridians; circular parallels:* that is, the angle between any two meridians is the same as their angular difference in longitude. This indicates a polar projection (azimuthal with the centre of the projection at one of the poles). Similar sub-categories as before can be identified:
  - *evenly spaced parallels:* a polar equidistant projection
  - *parallel spacing decreases away from the pole:* an azimuthal equal area projection
  - *parallel spacing increases away from the poles:* a polar stereographic projection.
- *Curved meridians:* this is more of a problem, as several types of projection fall into this category, including all oblique forms of azimuthal projection (except

gnomonic) as well as the transverse Mercator and many of the more esoteric projections. If the meridians are curved but all graticule intersections are right angles, it is more likely to be a conformal projection. Over a small area it may be difficult to determine whether, for example, the meridians are straight lines or are curved. For a typical 1 : 50 000 map sheet created on a transverse Mercator projection, for example, the curvature of a meridian amounts to the equivalent of around 1 mm off a straight line. However, large scale maps of a small area are less likely to present a problem because they are more susceptible to the kind of treatment outlined in the previous sections of this chapter.

Once the projection has been identified, the next stage is to determine the parameters that have been used, which can be done by taking measurements from the map. The key point to note here is that although it is not possible to determine all the parameters of the projection by taking measurements off the map, the ones that it is not possible to find are not usually necessary. For example, if no grid was shown on the map and the user in effect creates an arbitrary one, it is not possible to determine the correct false eastings and northings of the origin. On the other hand, the 'substitutes' that can be found are perfectly adequate if they are used consistently.

The most important parameters to be found are the ones that change the *shape* of the projection. These are detailed below for the main categories.

- *Normal cylindrical projections (contact at the equator):* As a general rule, there are no parameters that change the shape, as the projection is entirely defined once it is identified as conformal, equal area or equidistant. The exception to the rule is when a differential scaling has been applied to the meridians and parallels as in the case of the Peters projection (section 8.2). This can be identified by determining the scale factor in the two directions separately, as explained below.

- *Polar projections:* Again, the shape of the projection is entirely defined once it has been identified as stereographic, equidistant or equal area.

- *Conic projections:* The shape of the projection is affected by the choice of standard parallel (or parallels) and it is necessary to determine the value(s) used. Choose two widely spaced meridians that are apparent on the projection, which have a difference in longitude of $\Delta\lambda$, and measure the angle between them on the map, β. Then by a rearrangement of equation (10.1) the standard parallel, α, is given as

$$\alpha = \sin^{-1}\left(\frac{\beta}{\Delta\lambda}\right) \tag{12.14}$$

  If two standard parallels have been used, this value will be the average of the two. For conformal projections, any two that have a mean value α may be chosen without affecting the shape.

  This determines the origin of latitude. The origin of longitude is whichever meridian is aligned with grid north (whether a pre-existing grid or one that has been created for the purpose).

- *Transverse Mercator projection:* The shape of the projection is determined by the choice of central meridian. This may be found by measuring the convergence, γ, or the angle between a meridian and grid north. By measuring this angle at a point on the map whose latitude ϕ and longitude λ are known, the change in longitude $\Delta\lambda$ between this point and the central meridian can be found by rearranging equation (8.17) as

$$\Delta\lambda = \frac{\gamma}{\sin\phi} \tag{12.15}$$

  This gives the origin of longitude. The origin of latitude does not affect the shape, and therefore cannot be found. An arbitrary value can therefore be assigned.

If required, an overall rescaling can be found by comparing a distance on the map (for simplicity, one along a meridian) with the distance on the spherical model. Along a meridian, the distance on the sphere is

$$D_{\text{sphere}} = R\Delta\phi\frac{\pi}{180} \tag{12.16}$$

The latitude difference, $\Delta\phi$, is expressed in degrees; R is the Earth radius (a suitable value being 6371 km). This should be scaled by the appropriate map scale. The overall rescaling is then

$$k_0 = \frac{D_{\text{projection}}}{D_{\text{sphere}}} \tag{12.17}$$

This step is not actually necessary if the final step is for the map to be transformed into a new data set using common points, and the scale factor determined as part of this process. Finally, the false eastings and northings of the origin may be measured directly from the grid. An example of this procedure is given in the case study in section 13.5.

13
Case studies

13.1 Transformation of GPS data into a local datum

For this case study, data has been obtained by differential phase GPS observations, and will be transformed into the local datum using common points. Figure 13.1 shows the configuration of observations that were made, although the vectors have not been shown to avoid cluttering the diagram. A network adjustment of the observations indicates that the quality of the vectors is on average around 1.5 cm in plan and 2 cm in height. A reasonable spread of control points has been achieved: it was not possible

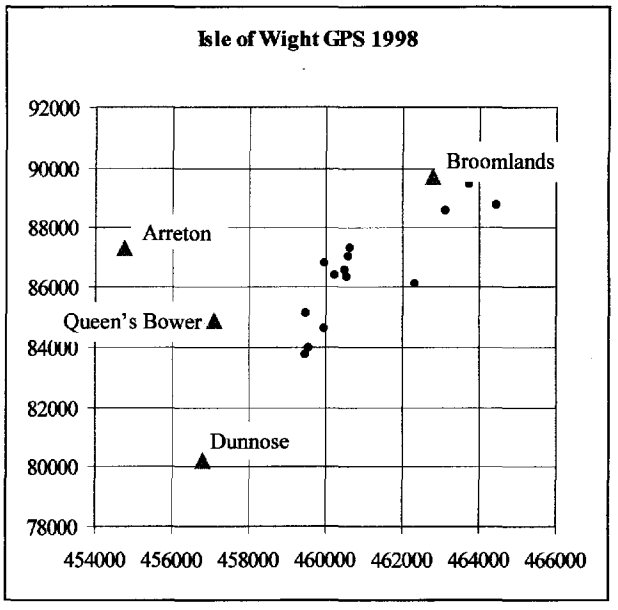

Figure 13.1 Configuration of observed points (Isle of Wight GPS 1998).

to surround the area completely as the south-east section of the map is in the sea. The

control points used are Ordnance Survey triangulation pillars. Several bench marks have also been included in the survey.

The initial transformation is a seven-parameter similarity transformation using the four control points shown in the diagram: Arreton, Broomlands, Dunnose, and Queen's Bower. The transformation parameters that result from this are shown in Table 13.1.

Table 13.1 Transformation parameters for seven-parameter similarity transformation

ΔX	−531.928 m
ΔY	167.961 m
ΔZ	−393.725 m
Rotation X	$-1.166''$
Rotation Y	$-5.115''$
Rotation Z	$1.382''$
Scale factor	+13.05 ppm

Note that

- The sizes of the translations are broadly in line with the transformations for the whole of the UK shown in section 4.2.2, allowing for a local distortion of the datum and the inaccuracy of the absolute GPS coordinates in WGS84.
- The rotations are not extreme.
- The scale factor (13 ppm) represents around 15 cm over this area, which is large, but not entirely unexpected for this datum.

The next step is to check the residuals of the transformation: that is, the difference between the original coordinates and the ones derived from the transformation. These are shown in Table 13.2.

Table 13.2 Residuals of the similarity transformation

	ϕ	λ	h
Arreton	−0.0077	−0.0154	0.0017
Broomlands	−0.0213	0.0098	0.0014
Dunnose	−0.0079	0.0235	0.0027
Queen's Bower	0.0368	−0.0179	−0.0057

As can be seen, these are generally below 2 cm, with the exception of the latitude residual at Queen's Bower, which is 3.7 cm. Mostly, these residuals are within the accuracy of the GPS survey. Before proceeding further, however, it is worth examining the extent to which an error in the control points would be noticeable by an examination of the residuals. To do this, an error is introduced into the eastings coordinate of Broomlands, increasing it by 20 cm. The residuals that result from this transformation are shown in Table 13.3.

The residuals are now rather larger in plan, typically around 5 cm, and this would perhaps be sufficient to alert the user. It is clear, however, that there is still not sufficient redundancy to isolate the problem to the coordinates of Broomlands.

Table 13.3 Residuals using erroneous coordinate set

	ϕ	λ	h
Arreton	−0.0589	−0.0471	0.0019
Broomlands	−0.0219	0.0614	0.0014
Dunnose	0.0430	0.0386	0.0016
Queen's Bower	0.0377	−0.0530	−0.0049

The necessity for redundancy is further illustrated by examining a situation with only three control points: in this case, eliminating Queen's Bower. The residuals for the transformation using the correct coordinates are shown in Table 13.4.

Table 13.4 Residuals using correct coordinate set and only three points

	ϕ	λ	h
Arreton	0.0064	−0.0198	0.0000
Broomlands	−0.0125	0.0055	0.0002
Dunnose	0.0061	0.0143	−0.0002

The residuals here are smaller in plan, and much smaller in height. Clearly, as neither the GPS data nor the control point coordinates have changed, this is a function of the reduced redundancy. Using the incorrect data set, the residuals shown in Table 13.5 are obtained.

Table 13.5 Residuals using incorrect coordinate set and only three points

	ϕ	λ	h
Arreton	0.0323	−0.0646	−0.0261
Broomlands	0.0110	0.0485	−0.0083
Dunnose	−0.0433	0.0161	0.0345

The residuals are larger, though not immediately registering a problem. The best indication that something is amiss with this data comes from the fact that the scale factor has now increased to around 24 ppm, although there is still no way of locating the source of the problem. The moral here is to use as many control points as possible, and to check whether the derived parameters are sensible numbers.

Returning to the original set of transformation parameters derived from the four points and the correct data, these are now applied to the entire data set to determine coordinates in the local datum. In this particular case, we have the advantage of having observed several bench marks in the area, and so it is now possible to compare the heights derived by the transformation with the known values. The comparison is shown in Fig. 13.2, with the residuals representing the known heights minus the transformed ones.

The residual values of the geoid (which are not actual geoid values, but what remains after the shifts and rotations) do seem to exhibit a pattern, as they are almost all negative and increasing from the north-east to the south-west before coming back to fit again at Dunnose. In fact, the sizes of residuals are not entirely inconsistent with errors in the GPS data, but the supposition that they are caused by geoid effects is

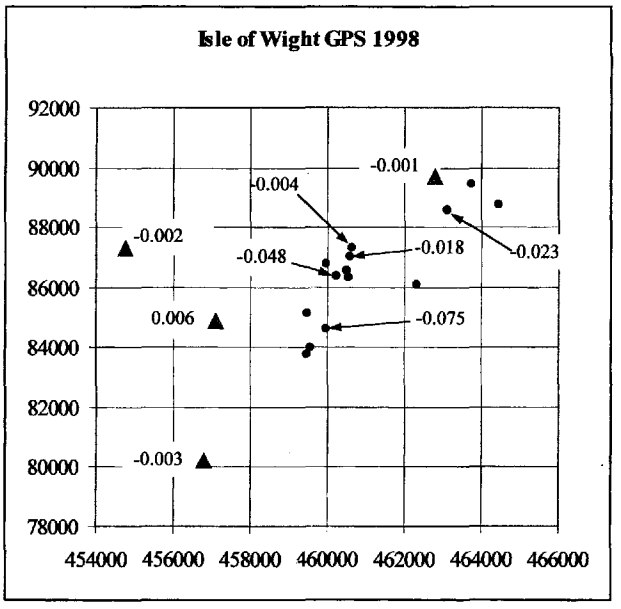

Figure 13.2 Residual geoid values after transformation (Isle of Wight GPS 1998).

supported by the fact that they would fit in with a geoid being 'drawn up' under the high ground to the south-west. A random pattern of residuals over this size of area would almost certainly indicate data errors.

A possible next step is therefore to transform the coordinates using an interpolation algorithm, as described in section 6.5. This essentially bypasses the classic approach of first converting the grid coordinates to cartesian via the procedures outlined in previous chapters, and instead stretches the GPS data to fit the grid coordinates and the heights as they are. This was carried out on the data set used in this example, and resulted in a mean scale factor of −339 ppm. The primary reason for such a large scale factor adjustment is the fact that no correction has been made for the projection. Therefore in a transformation of this type the real differences of scale between the two data sets (as found in the similarity transformation) have been masked by the much larger effect of the projection scale factor. Although this type of transformation can be used if no information is available on the projection, it does illustrate the advisability of first carrying out a similarity transformation if at all possible.

In most commercial packages, it is possible to choose the amount of distortion that is applied to the GPS data to make it fit onto the local control. In this instance, an average amount of distortion was chosen. With only four control points spread out widely across the area, it was not hard for the program to adjust the GPS data in two dimensions to fit the local control almost perfectly, resulting in very small residuals. In the height direction, the residuals are a function of the distance between bench marks and the amount of distortion required to fit the GPS data onto the given heights. The

largest height residuals are at those points where there is a large jump in the residual geoid distortion compared with other points nearby.

The parameters derived from this transformation were applied to the whole data set, and an alternative coordinate set derived for all points. As an illustration of the effect of the different transformations, the coordinates of one point in the centre of the network are shown in Table 13.6 for each transformation.

Table 13.6 Coordinates of Sandown Farm from the two transformations.

	Eastings	Northings	Height
Similarity transformation	459 467.603	85 163.762	18.294
Interpolation	459 467.585	85 163.821	18.371
Difference	0.018	−0.059	−0.077

Which is the correct answer? This question is almost impossible to answer, and this example simply illustrates the difficulty of finding coordinates in a distorted local datum to any better than a couple of centimetres. Instead, another pair of questions is posed:

- How important is it for the GPS results to be in sympathy with local mapping?
- How important is it for the GPS results to be easily integrated with new terrestrial observations?

In most cases the answer to the first question is 'just to a few centimetres', and the answer to the second question is 'quite important'. That is, an EDM traverse run between two of the new control points provided by GPS should not end up with large misclosures, and neither should runs of levelling.

The most appropriate solution in many cases is a hybrid of the similarity transformation and the interpolation, using the former for the plan and the latter for the height. This has the advantage of making the minimum change in the plan shape, thus allowing the new GPS control points to fit in well with new traverses. For height, the interpolation usually gives a better approximation of the distortions of the geoid, and is therefore more likely to fit in with new levelling. Some commercial packages allow this hybrid combination to be performed as a single operation.

13.2 A projection for Australia

13.2.1 The problem

Base maps of Australia are produced on a different projection or series of projections for each individual state. This is due to the fact that the original survey work had to be computed on projections that kept the scale factor distortion to a very low level. A company has acquired the data and wishes to put it all onto one projection, in order to avoid having discontinuities in the data. Area is not important, as this will be an attribute on the database, but features should 'look right'; that is, there should not be any great distortion of shape. Design a suitable projection.

largest height residuals are at those points where there is a large jump in the residual geoid distortion compared with other points nearby.

The parameters derived from this transformation were applied to the whole data set, and an alternative coordinate set derived for all points. As an illustration of the effect of the different transformations, the coordinates of one point in the centre of the network are shown in Table 13.6 for each transformation.

Table 13.6 Coordinates of Sandown Farm from the two transformations.

	Eastings	Northings	Height
Similarity transformation	459 467.603	85 163.762	18.294
Interpolation	459 467.585	85 163.821	18.371
Difference	0.018	−0.059	−0.077

Which is the correct answer? This question is almost impossible to answer, and this example simply illustrates the difficulty of finding coordinates in a distorted local datum to any better than a couple of centimetres. Instead, another pair of questions is posed:

- How important is it for the GPS results to be in sympathy with local mapping?
- How important is it for the GPS results to be easily integrated with new terrestrial observations?

In most cases the answer to the first question is 'just to a few centimetres', and the answer to the second question is 'quite important'. That is, an EDM traverse run between two of the new control points provided by GPS should not end up with large misclosures, and neither should runs of levelling.

The most appropriate solution in many cases is a hybrid of the similarity transformation and the interpolation, using the former for the plan and the latter for the height. This has the advantage of making the minimum change in the plan shape, thus allowing the new GPS control points to fit in well with new traverses. For height, the interpolation usually gives a better approximation of the distortions of the geoid, and is therefore more likely to fit in with new levelling. Some commercial packages allow this hybrid combination to be performed as a single operation.

13.2 A projection for Australia

13.2.1 The problem

Base maps of Australia are produced on a different projection or series of projections for each individual state. This is due to the fact that the original survey work had to be computed on projections that kept the scale factor distortion to a very low level. A company has acquired the data and wishes to put it all onto one projection, in order to avoid having discontinuities in the data. Area is not important, as this will be an attribute on the database, but features should 'look right'; that is, there should not be any great distortion of shape. Design a suitable projection.

13.2.2 A solution

If features are to 'look right', a conformal projection is called for in order to minimise the distortion of shape. In selecting a suitable developable surface, it should also be noted that the extent of the region to be mapped ranges in latitude from 45° S to 12° S, and in longitude from 110° E to 160° E. The country could therefore be described as a mid-latitude one with a broad spread in longitude, which leads to the suggestion of a conic projection. As it has already been decided that a conformal projection is required, the Lambert conformal conic projection is selected.

In selecting the required parameters, it should be remembered (see section 10.4) that the scale factor distortion is minimised when the standard parallels are selected as being $\frac{1}{6}$ of the range of latitude in from the extremes of the projection. Since the range of latitude is 37° (45° S – 12° S), this suggests 17° S and 39° S as suitable standard parallels. A plot of the scale factor as a function of latitude is shown in Fig. 13.3, on which the latitudes of the main population centres have been marked.

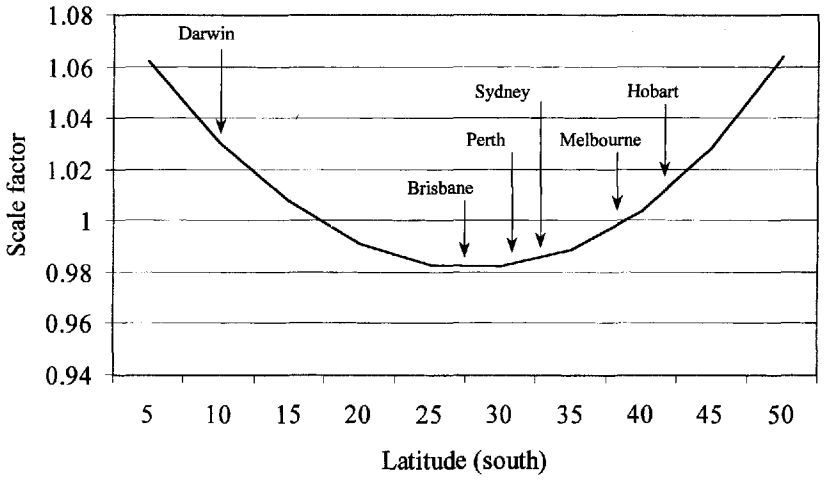

Figure 13.3 Scale factor for LCC with standard parallels at 17° S and 39° S.

It may considered more important to minimise scale factor distortion in the area of the main cities, an idea which is suggested by the fact that they happen to lie in a fairly narrow band of latitude. This could be done by selecting standard parallels at 25° S and 40° S, but this would of course lead to more extreme distortion on the edge of the projection.

With the parallels as they are, the minimum value of scale factor is 0.98 and the maximum is 1.03, which means a maximum distortion of 3%. Any areas measured from the map will be distorted by the square of this value, or 6% at the most.

The central meridian, or the origin of longitude, is selected as the mean of the extremes of longitude. This is 135° E. The maximum longitude difference from the central meridian is then 25°, which gives an indication of the convergence. From

section 10.1,

$$\gamma = \Delta\lambda \sin\alpha \tag{13.1}$$

where in this case α is the mean value of the two standard parallels, or 28° S. The maximum value of convergence is then around 12°, as shown in Fig. 13.4.

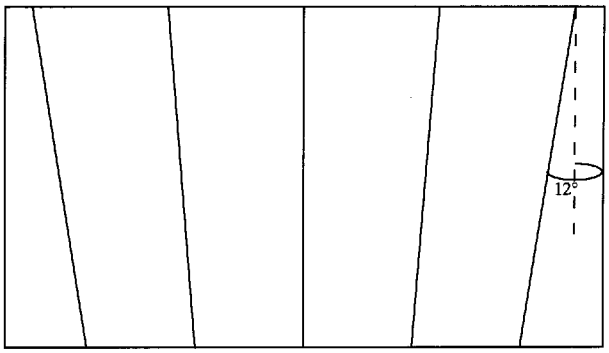

Figure 13.4 Appearance of the meridians on the selected LCC projection.

The final parameters to be selected are the origin of latitude and the values of the false coordinates. The choice of origin is not a critical decision, as it in no way affects the shape of the projection: for convenience ϕ_0 is then selected as 45° S, which gives all points in the region a positive value of northings. The false northings can then be selected as zero. Eastings are calculated as positive and negative quantities from the central meridian. A false eastings of +2500000 m would be sufficient to make all values positive.

13.3 Establishment of maritime boundaries on a projection

This case study will examine the problems that are encountered when using a projection to compute distances between points. The particular application that will be used is the establishment of maritime boundaries between states. A full description of the basis on which these boundaries are established is given in Beazley (1994). For the purposes of this example, it is sufficient to state that the main principle is the determination of the coordinates of points that are equidistant from the coastlines of different states.

This computation can be carried out exactly by using ellipsoidal coordinates. The algorithms used are complex, however, and not all GIS packages support their use. The process is made considerably easier if it is possible to use projection coordinates with no significant loss of accuracy. To begin with, a comment should be made about how not to do this. Some GIS packages allow the data to be displayed in what are termed 'geographical coordinates'. In reality this is nonsensical, as the very fact that the data are displayed on a flat screen implies that they are in a projection of some sort. This projection in fact appears as one in which the meridians and parallels are straight

lines with equal spacing, which was shown in section 8.1 as the cylindrical equidistant projection. In this case the scale factor is true along the meridians and equal to sec ϕ along the parallels. In a region such as the North Sea, at an average latitude of around 55° N, this translates as scale factors of 1.0 north–south and 1.74 east–west: this is certain to lead to some very wrong results.

A more appropriate approach is to design a projection specifically for use on this problem. Following similar arguments to those in section 13.2, a Lambert conformal conic projection is chosen with standard parallels at 51.5° N and 57.5° N. The scale factor for this projection varies within the area of interest between a minimum of 0.9986 and a maximum of 1.0017,as shown in Fig. 13.5. In presentational terms this is a very small variation, but the requirements for computation are rather more exacting.

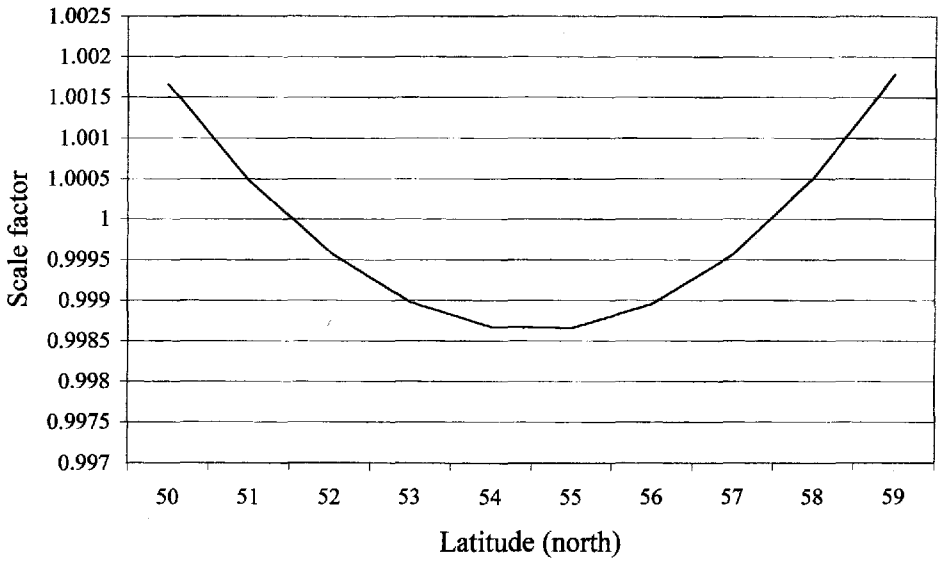

Figure 13.5 Scale factor in chosen projection (LCC).

Figure 13.6 shows a plot of the North Sea in the chosen projection, with an example point P that is proposed as being equidistant between A on the coastline of the UK and B on the coast of Norway. The scale factor at P in the direction of A is, by the definition of a conformal projection, equal to the scale factor in the direction of B. Unfortunately, this is not the case along the whole length of each line, as the scale factor will vary with latitude.

Between P and A the latitude changes from 56.5° to 54.5°, and therefore the scale factor variation is between 0.9992 and 0.9986. As a sufficient approximation the mean scale factor over the length of the line can be taken as the mean of these two values, or 0.9989. Between P and B the latitude change is from 56.5° to 58.5°, with the scale factor changing from 0.9992 to 1.0011. By a similar argument, the mean scale factor of the line is taken as 1.0002.

The approximate length of the lines PA and PB is 300 km. If this is the apparent length of the lines on the projection, Table 13.7 shows how the actual lengths of the lines differ.

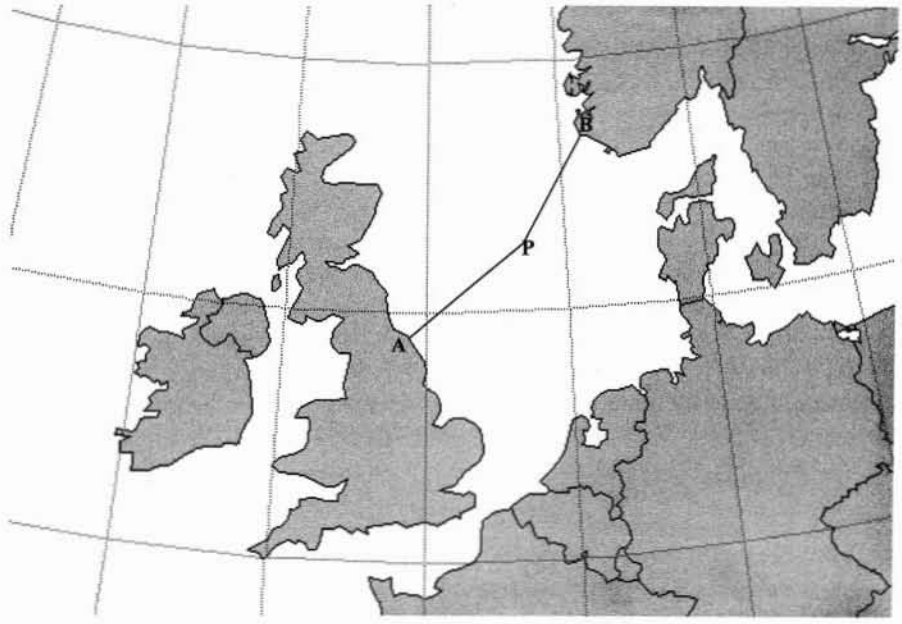

Figure 13.6 The North Sea in the chosen LCC projection.

Table 13.7 Actual and projected distances for sample point

Line	Projection distance (S km)	Scale factor (k)	Actual distance (S/k km)
PA	300	0.9989	300.330
PB	300	1.0002	299.940

Thus, two lines that appear equidistant on the projection actually differ by around 390 m. Although this level of accuracy may be sufficient for using the functionality of a GIS package to determine the most appropriate key points on the coast, it is unlikely to be sufficient for use in the final computation. In this case, it would be necessary to carry out the computations using ellipsoidal formulae. For smaller areas, it possible that a projection could be defined with sufficiently negligible variations to make the use of ellipsoidal formulae unnecessary.

13.4 Two-dimensional transformation of satellite imagery

The purpose of this case study is to examine the compatibility of two sets of data, and to see the extent to which the distortions due to the use of projections affects the accuracy that may be achieved through simple transformations.

In this example, a satellite image covering an area of around 60 km × 60 km is available for an area in the west of Wales. Several control points are visible on the image and can also be identified on a map of the area, and it is planned to carry out a two-dimensional transformation.

To examine the geometrical compatibility of the two data sets (the image and the map), the reasoning of section 12.1 is used. In that section, the concept of the scale factor ratio, r, was introduced. In this case, we are not really concerned with the projection distortions in the 'source' data set as, although a satellite image can be conceived of as being in a projection of a special kind, it covers such a small area that this effect is negligible. There will be distortions, but these will not be due to the projection effect. We are therefore really concerned only with the variations of scale factor in the underlying base map projection. Since the scale factor of the image is effectively 1, the scale factor of the map, k, will play the same role as the scale factor ratio, r.

The base map being used is an Ordnance Survey 1 : 25 000 map, and is therefore in the British National Grid, which is a transverse Mercator projection with the parameters given in section 8.4. This is illustrated in Fig. 13.7.

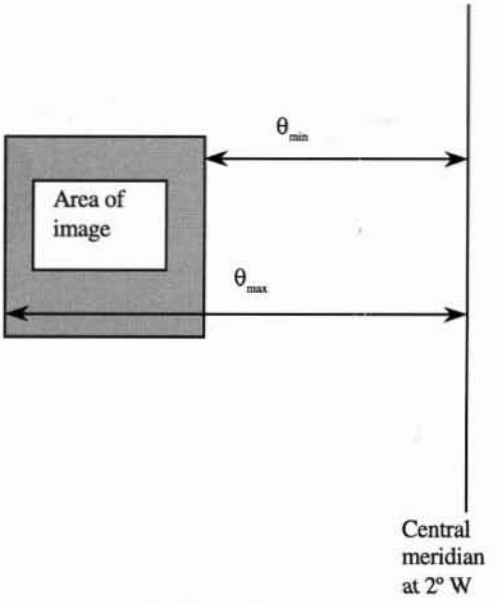

Figure 13.7 Position of the image in the projection.

The area concerned is at a latitude of 52° N and a longitude of 4° W, so the minimum value of the angular distance from the central meridian (θ_{min}) is around 1.23°. The other edge of the area concerned is around 60 km further away from the central meridian, or 0.72°, so θ_{max} is about 1.95°. These figures could be worked out without resorting to spherical trigonometry, by using the distances in kilometres and dividing by 6400 to express the solution in radians. The scale factor in this area therefore varies between a minimum of 0.999 83 and a maximum of 1.000 08, which have been computed using the secants of the maximum and minimum values of θ and an overall scaling of 0.9996. With a good spread of control points, applying a similarity transformation to the image would be the equivalent of using a mean overall scaling, in this case 0.999 95. Using the expression given in section 12.1 for the errors that are likely

to arise from the application of a similarity transformation gives

$$\varepsilon = 60000\left(\frac{1.00008}{0.99995} - 1\right) \tag{13.2}$$

which amounts to 7.4 m.

This is therefore not likely to be a problem in this situation, particularly if the pixel size is 10 m or more. In addition, if a similarity transformation does not lead to problems of geometrical incompatibility between the two data sets, there will certainly not be a problem when using the more complex transformations with multiple control points to overcome the other distortions that will be present in the image. In fact, this result could be said to be indicative of the general situation with satellite images: when transforming onto conformal projections over areas of a few tens of kilometres, and with pixel sizes of a few metres, there are unlikely to be problems resulting from the projection.

13.5 Two-dimensional transformation of GPS data

Some GPS software packages permit the transformation of data into a local reference system by a two-dimensional transformation, without having to enter the parameters of the projection used. This may be because they are not known, or it may just be for the convenience of the operator. It is not a procedure without risks, however, and this case study will explore some of the problems. In theoretical terms, this is exactly the same procedure as that carried out in the case study in section 13.4, except that the accuracy requirements are much higher.

Let us consider initially a survey that covers an area of 10 km^2, with the local control in a transverse Mercator projection, and the area under consideration a distance of 100 km from the central meridian. This is depicted in Fig. 13.8, with a set of four control points spread out around the area. The extremes of the scale factor can then be

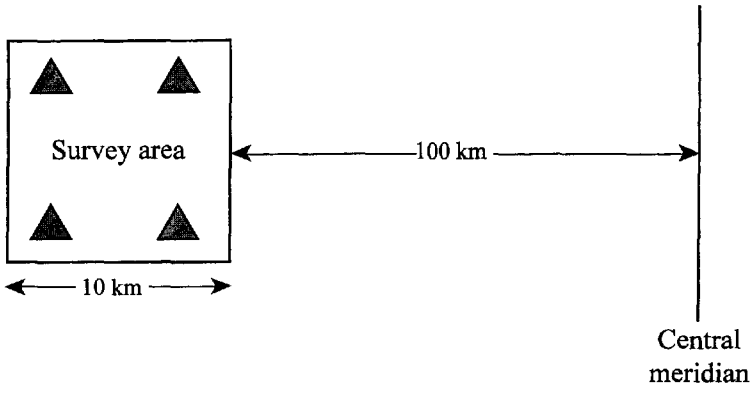

Figure 13.8 Position of survey with respect to the projection.

computed approximately as

$$k_{\min} = 0.9996 \sec\left(\frac{100}{6400}\right), \qquad k_{\max} = 0.9996 \sec\left(\frac{110}{6400}\right)$$

or 0.999 722 and 0.999 748 respectively. The mean scale factor that would be applied in a similarity transformation is 0.999 735. Then, by the argument outlined in section 12.1, the errors that would result from a similarity transformation are

$$\varepsilon = 10\,000\left(\frac{0.999\,75}{0.999\,735} - 1\right) = 0.15 \text{ m}$$

So, at 15 cm, this is really an unacceptable error when compared with the accuracy that can be obtained over this sort of distance with GPS.

Alternatively, if more control points were available across the region, a more complex warping of the GPS data could be carried out. Figure 13.9 shows a situation with an average spacing of 5 km between control points, a situation that is unlikely to be improved upon in practice.

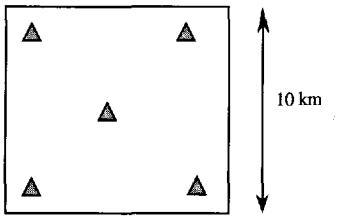

Figure 13.9 Additional control points used for warping transformation.

The scale factor between adjacent control points now ranges from 0.999 722 to 0.999 735, with a mean of 0.999 729. Applying equation (12.2) once again, this time with a distance of 5 km between points, gives

$$\varepsilon = 5000\left(\frac{0.999\,735}{0.999\,729} - 1\right) = 0.031$$

In other words, an error of 3 cm, which is again unlikely to be acceptable.

In conclusion, this type of transformation should be used with caution on surveys covering an area more than a couple of kilometres across. It is really suitable only for fitting GPS data onto a local site grid that is unrelated to a national or state coordinate system.

13.6 Determining the parameters of an unknown projection

This case study explores the situation where data from two different sources is to be combined, with at least one of the sources being a map with unknown projection parameters.

Let us assume, then, that data for the coastal areas around the British Isles is available, and is stored in the Mercator projection with known parameters. Data for the land areas has been digitised from a map that appears to be conic, but for which no other projection information is available. The size of the area used for this illustration is large, but this will assist the reader in visualising the processes at work. The appearance of the two data sets, as seen in Figs 13.10 and 13.11, is very different.

Initially, a similarity transformation is applied to the second data set, by matching points that can be identified in both maps. The results of this are shown in Fig. 13.12.

This is so far a poor attempt to combine the two, but a further improvement is next made by identifying many further common points and applying a rubber sheeting algorithm to pull the two together. The results of this stage are seen in Fig. 13.13.

This is an extremely unstable process for such radically different shapes, particularly away from the tie points, and for some applications it may be a problem that real changes in the coastline would be masked by the over-manipulation of the shape of the data set.

As an alternative, let us assume that the second projection is in fact a Lambert conformal conic (there is a slight increase in the distance between the parallels going north). The angle between two meridians with a difference in longitude $\Delta\lambda$ of 7.5° was measured as 5.45°. Then rearranging equation (10.1) gives

$$\sin\alpha = \frac{\gamma}{\Delta\lambda} = \frac{5.45}{7.5} \tag{13.3}$$

Hence the standard parallel can be inferred as being 46.5° N. Alternatively, this could be represented by two standard parallels an equal distance either side of this, say at 40° and 53° N.

These are the most fundamental parameters to be obtained, as they are the only ones that will have an effect on the *shape* of the map. The other parameters required can then be selected in a more or less arbitrary fashion. Having done this, it is then possible to convert the projection coordinates into geodetic ones, and then to re-project the data using the known parameters of the first projection.

The two data sets are again combined through a similarity transformation, and the results shown in Fig. 13.14. Although the match is still not perfect, this is a much more satisfactory starting point for applying more complex transformations through rubber sheeting: this will be a far more stable process, and real discrepancies in the two data sets will be far more apparent.

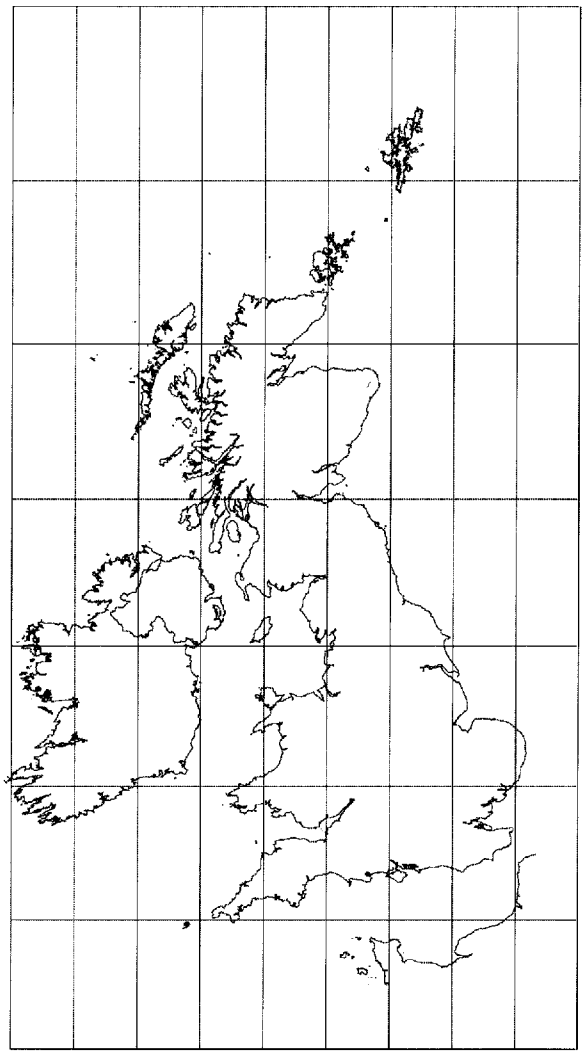

Figure 13.10 Data set in Mercator projection.

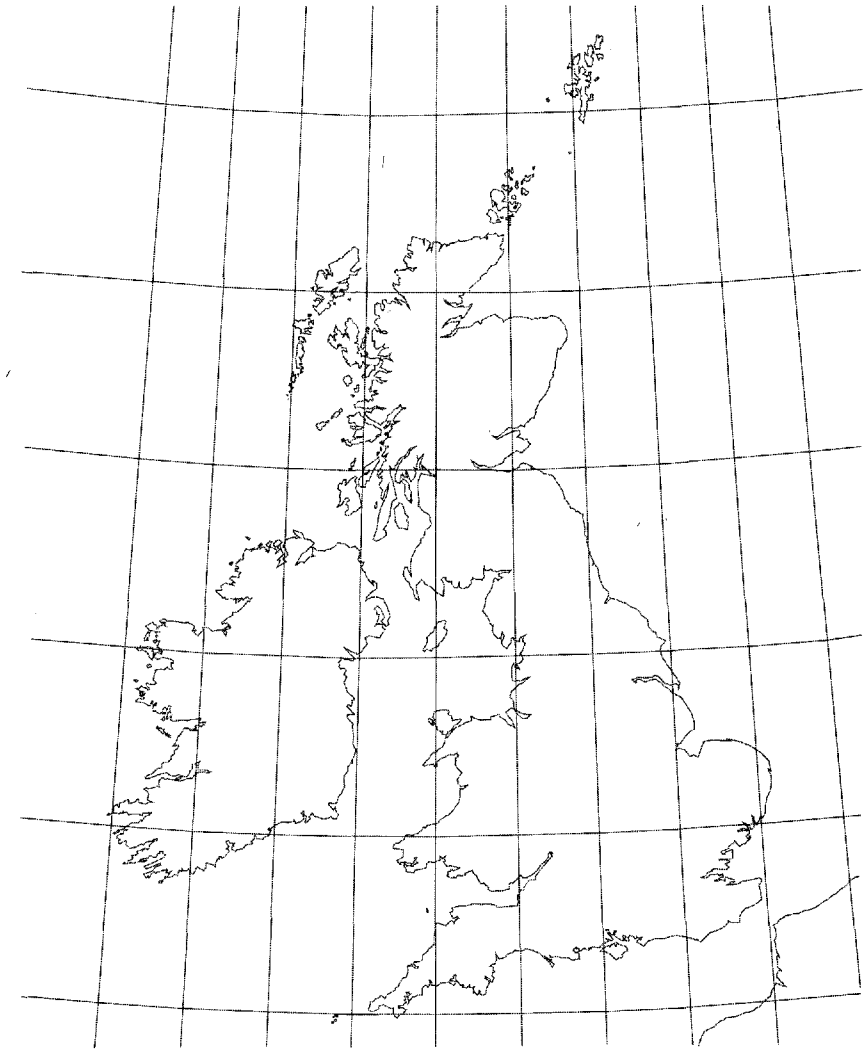

Figure 13.11 Data set in conic projection.

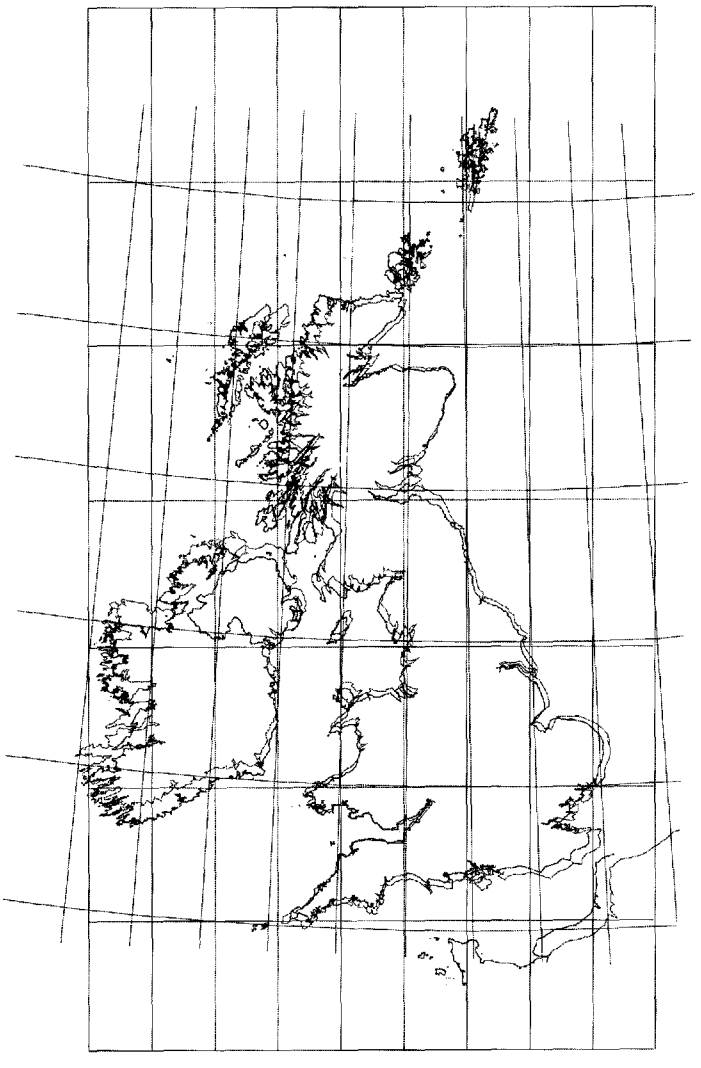

Figure 13.12 The two data sets combined with a similarity transformation.

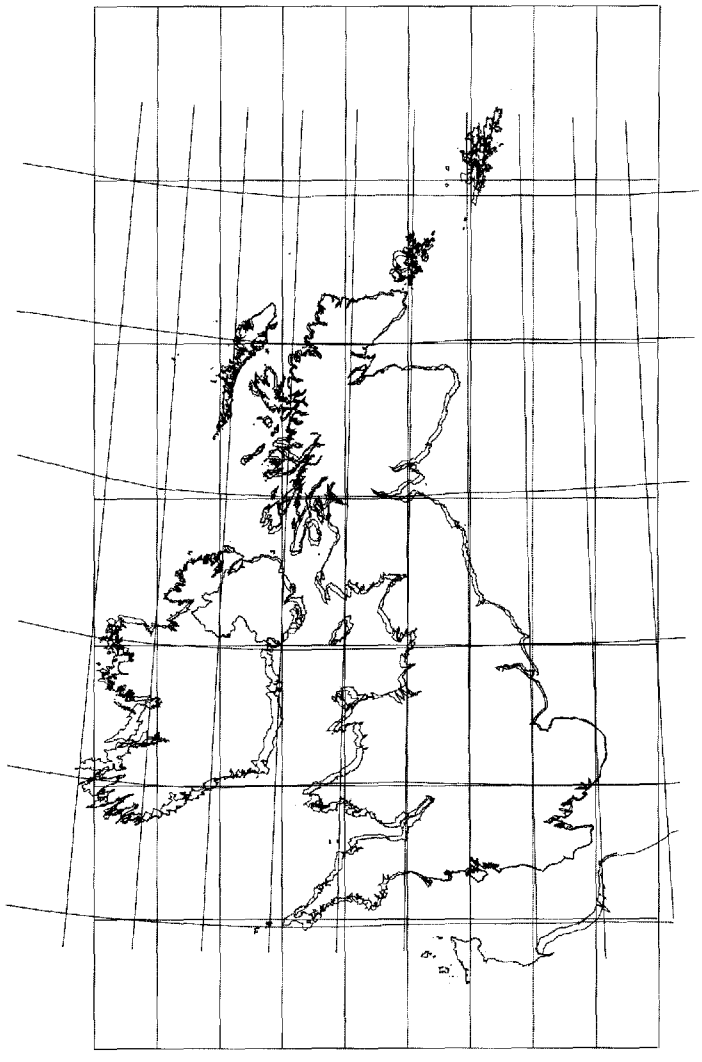

Figure 13.13 A further transformation is carried out by identifying common points and applying a rubber sheeting algorithm.

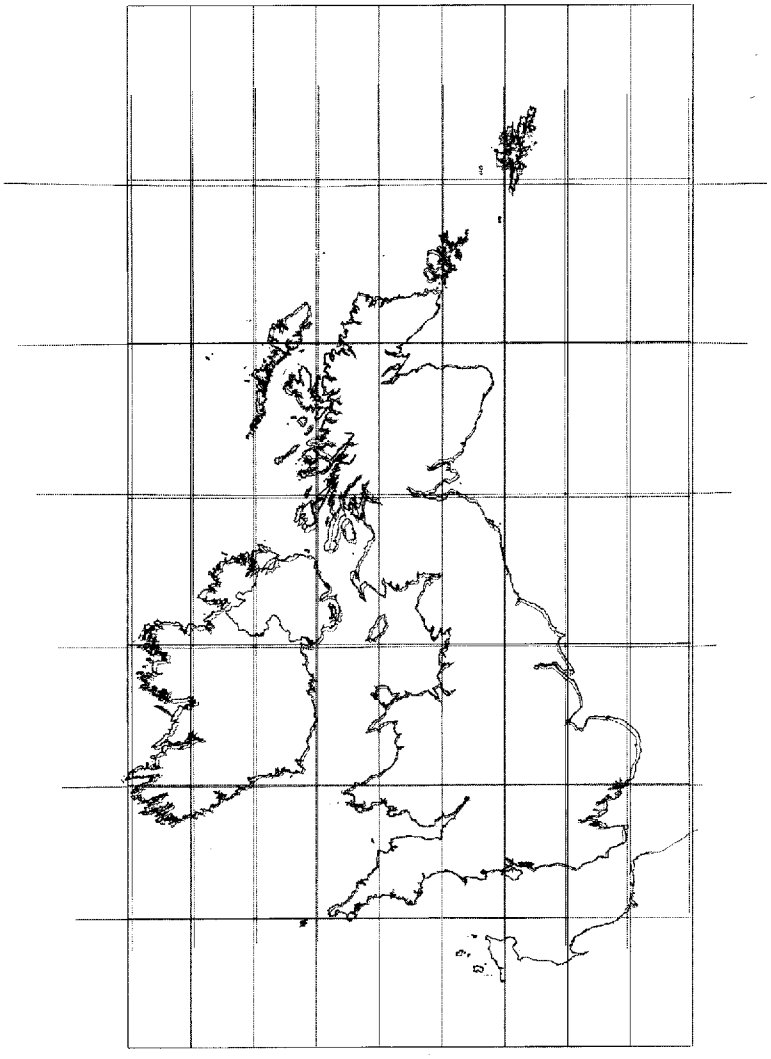

Figure 13.14 LCC map re-projected to the Mercator projection, and then transformed onto the original Mercator data via a similiarity transformation.

Appendix

A1 Spherical coordinates

This appendix is a summary of the formulae that are used to determine distances and azimuths from spherical coordinates.

The shortest route between two points on the sphere is referred to as the *great circle*. This can be defined by the intersection with the sphere of a plane that passes through both the points and the centre of the sphere.

Let the coordinates of two points, A and B, be given as (ϕ_A, λ_A) and (ϕ_B, λ_B) respectively. The great circle distance L_{AB} between the two points is given as

$$\cos L_{AB} = \sin\phi_A \sin\phi_B + \cos\phi_A \cos\phi_B \cos\Delta\lambda \tag{A1.1}$$

where $\Delta\lambda$ is the difference in longitude between the two points:

$$\Delta\lambda = \lambda_B - \lambda_A \tag{A1.2}$$

This gives an answer in angular units, which may be converted to distance by expressing the angle in radians and multiplying by an appropriate value for the radius of the spherical Earth:

$$L_{(\mathrm{km})} = 6371 \frac{\pi}{180} L_{(\mathrm{degrees})} \tag{A1.3}$$

The *azimuth* of the point B from A is defined as the clockwise angle between the meridian at A and the great circle, and is given by the expression

$$\cot A_{AB} = \frac{\cos\phi_A \tan\phi_B - \sin\phi_A \cos\Delta\lambda}{\sin\Delta\lambda} \tag{A1.4}$$

A2 Basic geometry of the ellipsoid

A2.1 Introduction

The ellipsoid is the basic reference surface for the definition of geodetic coordinates. With the use of GPS for most surveys that cover a large area, the need to carry out computations using geodetic coordinates has greatly diminished, and most local surveys will be carried out using projection coordinates that have been suitably corrected.

One of the few remaining uses for computations on the ellipsoid is in relation not to actual observations but to the establishment of boundaries: this may be, for example, between states or between oil concessions. Here it may be necessary, for example, to compute the coordinates of a boundary line between two defining points that lie several hundred kilometres apart. In such situations, the use of a projection is inappropriate, and the computations cannot be carried out in cartesian coordinates.

What follows is by no means a comprehensive discussion of all aspects of geometrical geodesy, but is a summary of some of the most useful concepts and formulae. A fuller treatment may be found in references such as Bomford (1980).

A2.2 Normal sections and geodesics

The azimuth from point A to point B (both having been projected onto the surface of the ellipsoid, if necessary) is defined firstly with reference to the plane that contains the normal to the ellipsoid at point A and both points A and B. The angle between this plane and the meridional section at A, when measured clockwise from north, defines the *normal section azimuth* from A to B, as shown in Fig. A.1.

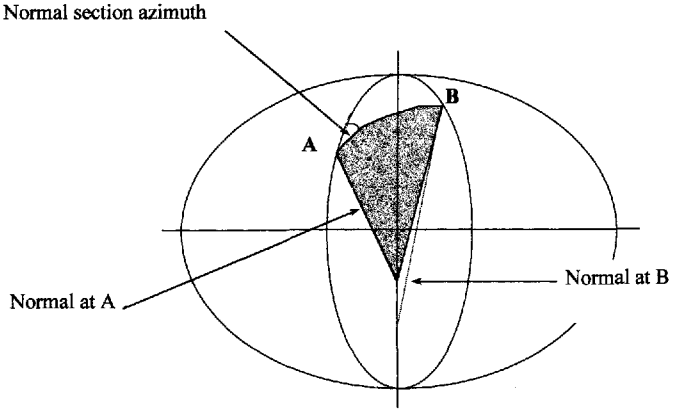

Figure A.1 Normal section at A to the point B.

The normal section from A to B is not, in general, the same plane as that from B to A. The intersection of these two planes with the surface of the ellipsoid therefore establishes two separate lines. Except in special cases, neither of these two lines is actually the shortest route between the two points.

The shortest route between two points on the surface of the spheroid is called the *geodesic*. It is a complex line that cannot be described by the intersection of a plane and the surface of the spheroid. As shown in Fig. A.2, it lies between the two normal sections with an initial azimuth at A that is closer to the normal section from A than the normal section from B in the ratio 2 : 1.

This property also applies in reverse, in that the azimuth of the geodesic at the point B will be closer to the normal section from B than to the normal section from A in the same ratio.

The maximum distance between the two normal sections is a function of the distance between the two points, and ranges from a few centimetres for lines up to 100 km

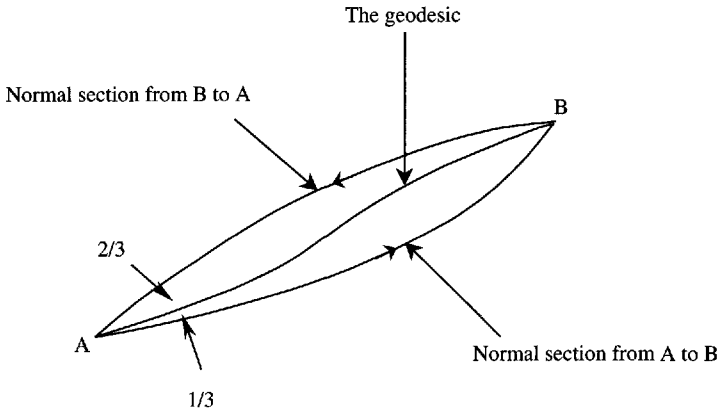

Figure A.2 Normal sections and the geodesic.

to several metres for lines up to 1000 km.

A2.3 Forward computation of coordinates

Several formulae may be used to determine the coordinates of a point given the coordinates of an initial point and the distance and azimuth. Some of them are less complex than the one quoted here, but have limited accuracy over longer distances.

Given the coordinates of point A (ϕ_A, λ_A) and a normal section azimuth (A) and distance (L), it is required to find the coordinates of the point B, such that:

$$\phi_B = \phi_A + \Delta\phi \tag{A2.1}$$

$$\lambda_B = \lambda_A + \Delta\lambda \tag{A2.2}$$

Clarke's 'best formulae' may be used for distances up to 1200 km with an accuracy of within $\frac{1}{25}$ of a part per million (Bomford, 1980). These formulae are:

$$\varepsilon = \frac{e^2}{1-e^2} \tag{A2.3}$$

$$r_2' = -\varepsilon \cos^2\phi_A \cos^2 A \tag{A2.4}$$

$$r_3' = 3\varepsilon(1-r_2')\cos\phi_A \sin\phi_A \cos A \tag{A2.5}$$

$$\theta = \frac{L}{\nu_A} - \frac{r_2'(1+r_2')}{6}\left(\frac{L}{\nu_A}\right)^3 - \frac{r_3'(1+3r_2')}{24}\left(\frac{L}{\nu_A}\right)^4 \tag{A2.6}$$

$$\left(\frac{\nu_A}{r}\right) = 1 - \frac{r_2'}{2}\theta^2 - \frac{r_3'}{6}\theta^3 \tag{A2.7}$$

$$\sin\psi = \sin\phi_A \cos\phi_A + \cos A \sin\theta \tag{A2.8}$$

Then

$$\sin\Delta\lambda = \sin A \sin\theta \sec\psi \tag{A2.9}$$

$$\tan\phi_B = (1+\varepsilon)\left\{1 - e^2\left(\frac{\nu_A}{r}\right)\frac{\sin\phi_A}{\sin\psi}\right\}\tan\psi \tag{A2.10}$$

All terms not specifically defined here are given in Chapter 2.

A2.4 Reverse computation of azimuth

A normal section azimuth, A_{AB}, between two points A (ϕ_A, λ_B) and B (ϕ_A, λ_B), which are of known coordinates, may be computed as follows:

$$\tan A_{AB} = \frac{-\Delta X \sin\lambda_A + \Delta Y \cos\lambda_A}{-\Delta X \sin\phi_A \cos\lambda_A - \Delta Y \sin\phi_A \sin\lambda_A + \Delta Z \cos\phi_A} \tag{A2.11}$$

The cartesian coordinates of A and B are found from the equations in section 2.4, and the convention for the differences is

$$\Delta X = X_B - X_A \tag{A2.12}$$

with similar terms for Y and Z.

A2.5 Determination of points on the geodesic

It is quite usual that a boundary between two areas (mineral concessions or neighbouring states) is defined as the geodesic between two points of given coordinates (on a certain datum). It will then be necessary to determine the coordinates of points at given intervals along the geodesic.

The fact that the geodesic is not a plane curve means that the methods for computing it are very complicated, usually involving the numerical expansion of an integration. Sharma (1966) gives a rigorous method for this. As an alternative, it is possible within a limited accuracy to compute points on the geodesic by interpolating between the normal sections. A suggested method follows.

1. Given the coordinates of the two end points, A and B, a distance L apart, compute the normal section azimuths from each end by the use of equations (A2.11) and (A2.12).

2. Compute the coordinates of a point on the normal section from A to B at a distance D using equations (A2.1)–(A2.10). Let this point be called P_A.

3. Compute the coordinates of a point on the normal section from B to A at a distance $(L - D)$, using the same formulae. Let this point be called P_B. The situation is illustrated in Fig. A.3.

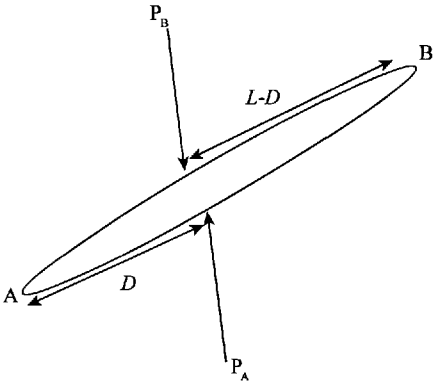

Figure A.3 Approximation of the geodesic.

It is now possible to interpolate the coordinates of the point on the geodesic from the coordinates of the two points found so far if we bear in mind that the geodesic is nearer to the normal section AB than to the section BA in the ratio 2 : 1 at the point A, halfway in between the two at the midpoint, and nearer BA than AB in the ratio 2 : 1 at the point B. Making an assumption of a linear change in the ratio, this gives the coordinates of the geodesic point (ϕ_G, λ_G) as

$$\phi_G = \phi_{P_A} + (\phi_{P_B} - \phi_{P_A}) \left(\frac{1 + (D/L)}{3} \right) \tag{A2.13}$$

$$\lambda_G = \lambda_{P_A} + (\lambda_{P_B} - \lambda_{P_A}) \left(\frac{1 + (D/L)}{3} \right) \tag{A2.14}$$

This technique has an estimated accuracy of within 10 cm at distances up to 1000 km.

A3 Determination of transformation parameters by least squares

A3.1 Introduction and least squares terminology

The purpose of this appendix is to bring together the mathematical processes for determining the parameters of any transformation model by least squares, given the coordinates in two different systems of a sufficient number of points. A basic familiarity with the least squares process has to be assumed, but the essential equations are recapitulated here to establish the terminology. The notation in this appendix follows that of Allan (1997b), and the specific application to transformation problems is based on the approach of Cooper (1987).

The aim of the least squares process is, in general, to determine best estimates of a set of *unobserved parameters*, **x**, from a set of *observed parameters*, **s**. The bold type indicates that we are dealing here with *vectors*, containing several parameters. For

this particular application, the unobserved parameters will usually be the parameters of the transformation, and the observed parameters will be coordinate values.

The relationship between $\mathbf{x}$ and $\mathbf{s}$ is expressed via a functional relationship:

$$F(\hat{\mathbf{x}};\hat{\mathbf{s}}) = 0 \tag{A3.1}$$

where $\hat{\mathbf{x}}$ represents the best estimate of $\mathbf{x}$ and $\hat{\mathbf{s}}$ the best estimate of $\mathbf{s}$.

Since in general we are dealing with non-linear transformations, the functional relationship is linearised as follows:

$$F(\hat{\mathbf{x}};\hat{\mathbf{s}}) = F(\dot{\mathbf{x}};\dot{\mathbf{s}}) + \frac{\partial F}{\partial \mathbf{X}}\mathbf{dx} + \frac{\partial F}{\partial \mathbf{s}}\mathbf{ds} \tag{A3.2}$$

where $\dot{\mathbf{x}}$ and $\dot{\mathbf{s}}$ represent provisional values of the unobserved and observed parameters respectively, and are related to the best estimates through the expressions

$$\hat{\mathbf{x}} = \dot{\mathbf{x}} + \mathbf{dx} \quad \text{and} \quad \hat{\mathbf{s}} = \dot{\mathbf{s}} + \mathbf{ds} \tag{A3.3}$$

In addition, the best estimates of the observed parameters are related to the observations themselves, $\overset{\circ}{\mathbf{s}}$, by the expression

$$\hat{\mathbf{s}} = \overset{\circ}{\mathbf{s}} + \mathbf{v} \tag{A3.4}$$

where the quantities $\mathbf{v}$ are described as the residuals.

Equations (A3.3) and (A3.4) can then be combined to express the vector $\mathbf{ds}$ in terms independent of the best estimates as

$$\mathbf{ds} = \overset{\circ}{\mathbf{s}} - \dot{\mathbf{s}} + \mathbf{v} \tag{A3.5}$$

$$= \mathbf{l} + \mathbf{v} \tag{A3.6}$$

In equation (A3.2), the functional relationship, F, will evaluate as zero when it is applied to the best estimates of the observed and the unobserved parameters, as this is the way that the relationship was first introduced in (A3.1). It is also the case that the relationship will be zero when applied to the provisional values of the observed and unobserved parameters, provided they are given as a consistent set. Thus, equation (A3.2) can be simplified to

$$\frac{\partial F}{\partial \mathbf{x}}\mathbf{dx} + \frac{\partial F}{\partial \mathbf{s}}\mathbf{ds} = 0 \tag{A3.7}$$

For convenience, $\mathbf{dx}$ will henceforward be referred to simply as $\mathbf{x}$, the vector of unknowns that must be added to the provisional values to obtain the best estimates. Similarly, the vector $\mathbf{ds}$ is refered to as $\mathbf{s}$, and is split into its component parts $\mathbf{l} + \mathbf{v}$. Thus the equation now becomes

$$\mathbf{Ax} + \mathbf{Cv} = -\mathbf{Cl} = \mathbf{b} \tag{A3.8}$$

where

$$\mathbf{A} = \frac{\partial F}{\partial \mathbf{x}} \tag{A3.9}$$

and

$$\mathbf{C} = \frac{\partial F}{\partial \mathbf{s}} \tag{A3.10}$$

A and **C** are termed *Jacobian matrices*, and represent the partial derivatives of the functional relationships with respect to the unobserved and observed parameters respectively. More concisely, they are often called the *design matrices*. Their form will be discussed in detail in relation to the different transformation models in the following sections, but it should be pointed out here that they have been derived from a linearisation that has the provisional values of the observed and unobserved parameters as its starting point. In evaluating the matrices, the provisional values of these parameters should therefore be used.

In some situations it is possible to express a problem in such a way that the design matrix **C** reduces to the identity matrix **I**. In such cases, equation (A3.8) can be rearranged and expressed in the form

$$\mathbf{Ax} = \mathbf{l} + \mathbf{v} \tag{A3.11}$$

If the observation equations are given in this form, it can be shown (Allan, 1997b) that the solution that minimises the weighted sum of the squares of the residuals, **v**, is

$$\mathbf{x} = (\mathbf{A}^{\mathrm{T}}\mathbf{WA})^{-1}(\mathbf{A}^{\mathrm{T}}\mathbf{Wl}) \tag{A3.12}$$

where **W** is an appropriate weight matrix.

Alternatively, if the more general form of equation (A3.8) is used, it can be shown (Allan, 1997b) that the solution now has the form

$$\mathbf{x} = \left(\mathbf{A}^{\mathrm{T}}(\mathbf{CW}^{-1}\mathbf{C}^{\mathrm{T}})^{-1}\mathbf{A}\right)^{-1}\left(\mathbf{A}^{\mathrm{T}}(\mathbf{CW}^{-1}\mathbf{C}^{\mathrm{T}})^{-1}\mathbf{b}\right) \tag{A3.13}$$

In the sections that follow, several different forms of transformation model will be considered. For each of these, the functional relationship between observed and unobserved parameters will first be defined, as will the weight matrix. The form of the design matrices will then be derived. After this, the solution proceeds by application of either equation (A3.12) or equation (A3.13).

Before proceeding to a consideration of the different transformation models, however, further consideration should be given to the formation of the vector **b** in (A3.8). The implication of the way that it was derived is that it is first necessary to form the vector **l** from a knowledge of the provisional values of the observed parameters. Although for many problems in least squares this is not hard, in the area of coordinate transformations it is often an unwieldy procedure. An alternative derivation of (A3.8) will therefore be given (as in Cross, 1983) that obviates the need to form provisional values of the observations.

This starts with the alternative definition

$$F(\hat{\mathbf{x}};\hat{\mathbf{s}}) = F(\mathring{\mathbf{x}};\mathring{\mathbf{s}}) + \frac{\partial F}{\partial \mathbf{x}}\mathbf{dx} + \frac{\partial F}{\partial \mathbf{s}}\mathbf{v} \tag{A3.14}$$

which follows from the fact that the residuals, **v**, are defined as the differences between the observations and the best estimates. This can then be simplified to

$$\mathbf{Ax} + \mathbf{Cv} = -F(\dot{\mathbf{x}}; \mathring{\mathbf{s}}) = \mathbf{b} \tag{A3.15}$$

Therefore **b** can be found by substituting the observations (observed coordinates) and the provisional values of the unobserved parameters (transformation parameters) into the functional relationship.

It should also be noted that, with the linearisation now starting from the provisional values of the unobserved parameters and the observed values of the observed parameters, it is these values that should be used in evaluating the matrices **A** and **C**. This is not always stated explicitly in the derivations that follow, as it would make the notation rather unwieldy, but it is implicit throughout.

A3.2 Two-dimensional transformations

A3.2.1 The similarity transformation

A set of coordinates in the system (x, y) is to be transformed into the system (X, Y). Consider the simple similarity transformation shown in Fig. A.4, in which the coordinates undergo a rotation θ, a scaling factor λ, and shifts of Δx and Δy in the axes.

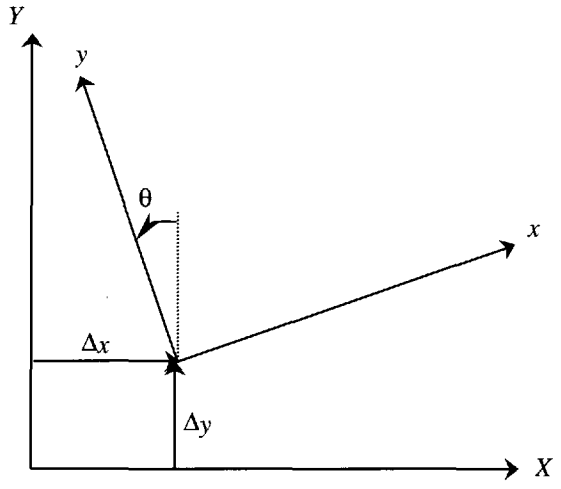

Figure A.4 Two-dimensional similarity transformation.

This transformation can be expressed by the following equations:

$$\begin{aligned} X &= x\lambda\cos\theta - y\lambda\sin\theta + \Delta x \\ Y &= x\lambda\sin\theta + y\lambda\cos\theta + \Delta y \end{aligned} \tag{A3.16}$$

The first step is to recast these expressions into the form of a set of functional relationships in the form required by equation (A3.1). At the same time, the terms in λ and θ

may be re-expressed as two new parameters, a and b. Thus

$$F_1 = ax - by + \Delta x - X \qquad \text{(A3.17)}$$
$$F_2 = bx + ay + \Delta y - Y$$

where

$$a = \lambda\cos\theta, \qquad b = \lambda\sin\theta \qquad \text{(A3.18)}$$

In solving for a and b, the original parameters λ and θ can subsequently be recovered from

$$\lambda = \sqrt{a^2 + b^2} \qquad \text{(A3.19)}$$

$$\theta = \tan^{-1}\frac{b}{a} \qquad \text{(A3.20)}$$

The situation is further simplified if it is possible to say that the coordinates (x,y) are effectively constants, and may be regarded as being without error. In this case, the only observed parameters are the coordinates (X,Y), and the unobserved parameters are a, b, Δx and Δy.

The design matrices **A** and **C** then have the form

$$\mathbf{A} = \begin{pmatrix} \frac{\partial F_1}{\partial a} & \frac{\partial F_1}{\partial b} & \frac{\partial F_1}{\partial \Delta x} & \frac{\partial F_1}{\partial \Delta y} \\ \frac{\partial F_2}{\partial a} & \frac{\partial F_2}{\partial b} & \frac{\partial F_2}{\partial \Delta x} & \frac{\partial F_2}{\partial \Delta y} \end{pmatrix} \qquad \text{(A3.21)}$$

$$\mathbf{C} = \begin{pmatrix} \frac{\partial F_1}{\partial X} & \frac{\partial F_1}{\partial Y} \\ \frac{\partial F_2}{\partial X} & \frac{\partial F_2}{\partial Y} \end{pmatrix} \qquad \text{(A3.22)}$$

These are then evaluated as:

$$\mathbf{A} = \begin{pmatrix} x & -y & 1 & 0 \\ x & y & 0 & 1 \end{pmatrix} \qquad \text{(A3.23)}$$

$$\mathbf{C} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \text{(A3.24)}$$

As the matrix **C** found above has the form $-\mathbf{I}$, it can be seen that the general form of the observation equations given in (A3.8) can be reduced to the special case of equation (A3.11).

In fact, the matrices determined in (A3.23) and (A3.24) are only a part of the full picture, as these have been derived for only one point. In general, there will be a similar matrix for each point that is common to the two coordinate systems. Denoting the partial matrix by the symbol $\mathbf{A}_i$, where

$$\mathbf{A}_i = \begin{pmatrix} x_i & -y_i & 1 & 0 \\ x_i & y_i & 0 & 1 \end{pmatrix} \qquad \text{(A3.25)}$$

and (x_i, y_i) are the coordinates of the ith point, the full design matrix for n points is made up as follows:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_n \end{pmatrix} \tag{A3.26}$$

which is a $(2n \times 4)$ matrix.

The more general form of the transformation is the situation where it is not possible to assume that one set of coordinates is correct, and instead both (X, Y) and (x, y) are assumed to be observed data with associated weights. The functional relationships are the same as before, but they will now be written in a form that makes it explicit that they are two equations out of a total set of $2n$. Thus:

$$F_{1i} = ax_i - by_i + \Delta x - X_i$$
$$F_{2i} = bx_i + ay_i + \Delta y - Y_i$$

As before, the sub-matrix $\mathbf{A}_i$ is determined by differentiating with respect to the unobserved parameters, and is exactly the same as previously:

$$\mathbf{A}_i = \begin{pmatrix} x_i & -y_i & 1 & 0 \\ x_i & y_i & 0 & 1 \end{pmatrix} \tag{A3.28}$$

The sub-matrix $\mathbf{C}_i$ is now different, however, as it includes the derivatives with respect to all observed parameters, thus:

$$\mathbf{C}_i = \begin{pmatrix} \dfrac{\partial F_{1i}}{\partial x_{1i}} & \dfrac{\partial F_{1i}}{\partial y_{1i}} & \dfrac{\partial F_{1i}}{\partial X_{1i}} & \dfrac{\partial F_{1i}}{\partial Y_{1i}} \\ \dfrac{\partial F_{1i}}{\partial x_{2i}} & \dfrac{\partial F_{1i}}{\partial y_{2i}} & \dfrac{\partial F_{1i}}{\partial X_{2i}} & \dfrac{\partial F_{1i}}{\partial Y_{2i}} \end{pmatrix} \tag{A3.29}$$

which can be evaluated as

$$\mathbf{C}_i = \begin{pmatrix} a & -b & -1 & 0 \\ b & a & 0 & -1 \end{pmatrix} \tag{A3.30}$$

The elements of the misclosure vector are determined from

$$\mathbf{b}_i = -\begin{pmatrix} F_{1i}(\dot{\mathbf{x}}, \mathring{\mathbf{s}}) \\ F_{2i}(\dot{\mathbf{x}}, \mathring{\mathbf{s}}) \end{pmatrix} = \begin{pmatrix} ax_i - by_i + \Delta x - X_i \\ bx_i - ay_i + \Delta y - Y_i \end{pmatrix} \tag{A3.31}$$

The full set of equations now fits together as follows:

$$\begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_n \end{pmatrix} \begin{pmatrix} x_a \\ x_b \\ x_{\Delta x} \\ x_{\Delta y} \end{pmatrix} + \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_n \end{pmatrix} \begin{pmatrix} \mathbf{v}_{x_1} \\ \mathbf{v}_{y_1} \\ \mathbf{v}_{X_1} \\ \mathbf{v}_{Y_1} \\ \mathbf{v}_{x_2} \\ \vdots \\ \mathbf{v}_{Y_n} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{pmatrix} \tag{A3.32}$$

The dimensions of these matrices are

$$(2n \times 4)(4 \times 1) + (2n \times 4n)(4n \times 1) = (2n \times 1)$$

Note that each **0** in the **C** matrix is a sub-matrix of dimensions (2×4).

It is helpful to recall here that the equations derived above should be evaluated using the provisional values of the unobserved parameters (the transformation parameters) and the observed values of the observations (the coordinate sets). The equations could be made more explicit by writing terms such as

$$\dot{a}, \dot{b}, \Delta\dot{x}, \text{ and } \Delta\dot{y}$$

for the transformation parameters, and

$$\mathring{x}_i, \mathring{y}_i, \mathring{X}_i, \text{ and } \mathring{Y}_i$$

for the observed parameters; in practice, little confusion can arise as these will in general be the only values available.

Finally, to obtain a solution through the use of equation (A3.13), it is necessary to introduce an appropriate weight matrix. The most important point about this is that it must be consistent with the order of the observed parameters implied by the formation of the **C** matrix and the **v** vector. Thus, the weight matrix **W** would take the form

$$\mathbf{W}^{-1} = \begin{pmatrix} \sigma_{x1}^2 & \sigma_{x_1y_1} & 0 & 0 & \sigma_{x_1x_2} & \cdots & \cdots \\ \sigma_{x_1y_1} & \sigma_{y1}^2 & 0 & 0 & \sigma_{y_1x_2} & \cdots & \cdots \\ 0 & 0 & \sigma_{X_1}^2 & \sigma_{X_1Y_1} & 0 & \cdots & \cdots \\ 0 & 0 & \sigma_{X_1Y_1} & \sigma_{Y_1}^2 & 0 & \cdots & \cdots \\ \sigma_{x_1x_2} & \sigma_{y_1x_2} & 0 & 0 & \sigma_{x_2}^2 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \sigma_{Y_n}^2 \end{pmatrix} \tag{A3.33}$$

where $\sigma_{x_1}^2$ is the variance of the observed coordinate of x_1, and $\sigma_{x_1y_1}$ is the covariance between coordinates x_1 and y_1.

This has been formed under the assumption that the two data sets are independent of each other, and thus that all the covariances between the two sets are zero.

A3.2.2 Affine transformation

The affine transformation was introduced in Chapter 12. The basic equations have the form

$$X = a_0 + a_1x + a_2y \tag{A3.34}$$

$$Y = b_0 + b_1x + b_2y \tag{A3.35}$$

The functional relationships are therefore defined as

$$F_{1i} = a_0 + a_1 x_i + a_2 y_i - X_i \tag{A3.36}$$

$$F_{2i} = b_0 + b_1 x_i + b_2 y_i - Y_i \tag{A3.37}$$

The sub-elements of the design matrices are

$$\mathbf{A}_i = \left[\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right] \begin{pmatrix} 1 & x_i & y_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_i & y_i \end{pmatrix} \tag{A3.38}$$

$$\mathbf{C}_i = \left[\frac{\partial \mathbf{F}}{\partial \mathbf{s}}\right] \begin{pmatrix} a_1 & a_2 & -1 & 0 \\ b_1 & b_2 & 0 & -1 \end{pmatrix} \tag{A3.39}$$

and the misclosure vector is

$$\mathbf{b}_i = -\begin{pmatrix} F_{1i}(\dot{\mathbf{x}}, \mathring{\mathbf{s}}) \\ F_{2i}(\dot{\mathbf{x}}, \mathring{\mathbf{s}}) \end{pmatrix} = -\begin{pmatrix} a_0 + a_1 x_i + a_2 y_i - X_i \\ b_0 + b_1 x_i + b_2 y_i - Y_i \end{pmatrix} \tag{A3.40}$$

These elements fit together as

$$\begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_n \end{pmatrix} \begin{pmatrix} x_{a_0} \\ x_{a_1} \\ x_{a_2} \\ x_{b_0} \\ x_{b_1} \\ x_{b_2} \end{pmatrix} + \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_n \end{pmatrix} \begin{pmatrix} \mathbf{v}_{x_1} \\ \mathbf{v}_{y_1} \\ \mathbf{v}_{X_1} \\ \mathbf{v}_{Y_1} \\ \mathbf{v}_{x_2} \\ \vdots \\ \mathbf{v}_{Y_n} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{pmatrix} \tag{A3.41}$$

The dimensions of the matrices are

$$(2n \times 6)(6 \times 1) + (2n \times 4n)(4n \times 1) = (2n \times 1)$$

Each $\mathbf{0}$ in $\mathbf{C}$ is again a (2×4) sub-matrix. The weight matrix is again as in (A3.33).

A3.2.3 Second-order polynomials

The equations for the two-dimensional transformation by second- order polynomials were introduced in Chapter 12 as

$$X = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 y^2 + a_5 xy \tag{A3.42}$$

$$Y = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 y^2 + b_5 xy \tag{A3.43}$$

The functional relationships follow from these definitions:

$$F_{1i} = a_0 + a_1 x_i + a_2 y_i + a_3 x_i^2 + a_4 y_i^2 + a_5 x_i y_i - X_i \tag{A3.44}$$

$$F_{2i} = b_0 + b_1 x_i + b_2 y_i + b_3 x_i^2 + b_4 y_i^2 + b_5 x_i y_i - Y_i \tag{A3.45}$$

The sub-elements of the design matrices and the misclosure vector are

$$\mathbf{A}_i = \left[\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right] = \begin{pmatrix} 1 & x_i & y_i & x_i^2 & y_i^2 & x_i y_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & x_i & y_i & x_i^2 & y_i^2 & x_i y_i \end{pmatrix} \tag{A3.46}$$

$$\mathbf{C}_i = \left[\frac{\partial \mathbf{F}}{\partial \mathbf{s}}\right] = \begin{pmatrix} a_1 + 2a_3 x_i + a_5 y_i & a_2 + 2a_4 y_i + a_5 x_i & -1 & 0 \\ b_1 + 2b_3 x_i + b_5 y_i & b_2 + 2b_4 y_i + b_5 x_i & 0 & -1 \end{pmatrix} \tag{A3.47}$$

$$\mathbf{b}_i = -\begin{pmatrix} F_{1i}(\dot{\mathbf{x}}, \mathring{\mathbf{s}}) \\ F_{2i}(\dot{\mathbf{x}}, \mathring{\mathbf{s}}) \end{pmatrix} = -\begin{pmatrix} a_0 + a_1 x_i + a_2 y_i + a_3 x_i^2 + a_4 y_i^2 + a_5 x_i y_i - X_i \\ b_0 + b_1 x_i + b_2 y_i + b_3 x_i^2 + b_4 y_i^2 + b_5 x_i y_i - Y_i \end{pmatrix} \tag{A3.48}$$

The full equations are then constructed as in equation (A3.41), except that the vector of unobserved parameters is

$$\mathbf{x}^{\mathsf{T}} = \begin{pmatrix} x_{a_0} & x_{a_1} & x_{a_2} & \cdots & x_{b_4} & x_{b_5} \end{pmatrix} \tag{A3.49}$$

The dimensions of the matrices are

$$(2n \times 12)(12 \times 1) + (2n \times 4n)(4n \times 1) = (2n \times 1)$$

A3.3 Three-dimensional transformations

A3.3.1 The seven-parameter similarity transformation

The seven-parameter three-dimensional similarity transformation was introduced in Chapter 4 in the form

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} + (1 + \lambda)\mathbf{R} \begin{pmatrix} x \\ y \\ x \end{pmatrix} \tag{A3.50}$$

where $\mathbf{R}$ is the rotation matrix given by

$$\mathbf{R} = \begin{pmatrix} 1 & \alpha_3 & -\alpha_2 \\ -\alpha_3 & 1 & \alpha_1 \\ \alpha_2 & -\alpha_1 & 1 \end{pmatrix} \tag{A3.51}$$

and α_1, α_2, and α_3 are rotations about the x, y, and z axes respectively. The functional relationships follow from the above definition, and are conveniently expressed in matrix form as

$$\mathbf{F}_i = \Delta\mathbf{X} + (1 + \lambda)\mathbf{R}\mathbf{x} - \mathbf{X} \tag{A3.52}$$

where

$$\Delta\mathbf{X} = \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \qquad \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{A3.53}$$

The design matrix **A** is then developed as follows:

$$\mathbf{A}_i = \left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right) = \left(\frac{\partial \mathbf{F}}{\partial \lambda} \quad \frac{\partial \mathbf{F}}{\partial \alpha_1} \quad \frac{\partial \mathbf{F}}{\partial \alpha_2} \quad \frac{\partial \mathbf{F}}{\partial \alpha_3} \quad \frac{\partial \mathbf{F}}{\partial \Delta \mathbf{X}}\right) \tag{A3.54}$$

$$= \left(\mathbf{Rx} \quad (1+\lambda)\left(\frac{\partial \mathbf{R}}{\partial \alpha_1}\right)\mathbf{x} \quad (1+\lambda)\left(\frac{\partial \mathbf{R}}{\partial \alpha_2}\right)\mathbf{x} \quad (1+\lambda)\left(\frac{\partial \mathbf{R}}{\partial \alpha_3}\right)\mathbf{x} \quad \mathbf{I}\right)$$

The first element in equation (A3.53) can be evaluated as

$$\mathbf{Rx} = \begin{pmatrix} 1 & \alpha_3 & -\alpha_2 \\ -\alpha_3 & 1 & \alpha_1 \\ \alpha_2 & -\alpha_1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x+\alpha_3 y-\alpha_2 z \\ -\alpha_3 x+y+\alpha_1 z \\ \alpha_2 x-\alpha_1 y+z \end{pmatrix} \tag{A3.54}$$

The most common application of this transformation is to the conversion of coordinates from one geodetic datum to another. In these situations, the rotations will be very small (a few seconds of arc) and it will be the case that

$$x \gg \alpha_3 y \tag{A3.56}$$

and similar simplifications can therefore reduce (A3.55) to

$$\mathbf{Rx} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{A3.57}$$

The other terms in (A3.54) are evaluated as

$$(1+\lambda)\left(\frac{\partial \mathbf{R}}{\partial \alpha_1}\right)\mathbf{x} = (1+\lambda)\mathbf{R}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ x \end{pmatrix} = (1+\lambda)\mathbf{R}\begin{pmatrix} 0 \\ z \\ -y \end{pmatrix}$$

$$= (1+\lambda)\begin{pmatrix} \alpha_3 z+\alpha_2 y \\ z-\alpha_1 y \\ \alpha_1 z-y \end{pmatrix} \approx (1+\lambda)\begin{pmatrix} 0 \\ z \\ -y \end{pmatrix} \approx \begin{pmatrix} 0 \\ z \\ -y \end{pmatrix} \tag{A3.58}$$

$$(1+\lambda)\left(\frac{\partial \mathbf{R}}{\partial \alpha_2}\right)\mathbf{x} \approx \begin{pmatrix} -z \\ 0 \\ x \end{pmatrix} \tag{A3.59}$$

$$(1+\lambda)\left(\frac{\partial \mathbf{R}}{\partial \alpha_3}\right)\mathbf{x} \approx \begin{pmatrix} y \\ -x \\ 0 \end{pmatrix} \tag{A3.60}$$

The complete set of equations forms the sub-element of the design matrix **A** thus:

$$\mathbf{A}_i = \begin{pmatrix} x & 0 & -z & y & 1 & 0 & 0 \\ y & z & 0 & -x & 0 & 1 & 0 \\ z & -y & x & 0 & 0 & 0 & 1 \end{pmatrix} \tag{A3.61}$$

The sub-element of the design matrix $\mathbf{C}$ is formed from

$$\mathbf{C}_i = \left[\frac{\partial \mathbf{F}}{\partial \mathbf{s}}\right] = \begin{pmatrix}(1+\lambda)\mathbf{R} & -\mathbf{I}\end{pmatrix} \approx \begin{pmatrix} 1 & \alpha_3 & -\alpha_2 & -1 & 0 & 0 \\ -\alpha_3 & 1 & \alpha_1 & 0 & -1 & 0 \\ \alpha_2 & -\alpha_1 & 1 & 0 & 0 & -1 \end{pmatrix} \tag{A3.62}$$

The misclosure vector is given by

$$\mathbf{b}_i = -\mathbf{F}_i(\mathring{\mathbf{x}}, \mathring{\mathbf{s}}) = \Delta\mathbf{X} + (1+\lambda)\mathbf{R}\mathbf{x} - \mathbf{X} \tag{A3.63}$$

The matrix sub-elements combine in the full set of equations as

$$\begin{pmatrix}\mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_n\end{pmatrix}\begin{pmatrix} x_\lambda \\ x_{\alpha_1} \\ x_{\alpha_2} \\ x_{\alpha_3} \\ x_{\Delta X} \\ x_{\Delta Y} \\ x_{\Delta Z}\end{pmatrix} + \begin{pmatrix}\mathbf{C}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_n\end{pmatrix}\begin{pmatrix}\mathbf{v}_{X_1} \\ \mathbf{v}_{Y_1} \\ \mathbf{v}_{X_1} \\ \mathbf{v}_{Y_1} \\ \mathbf{v}_{X_2} \\ \vdots \\ \mathbf{v}_{Y_n}\end{pmatrix} = \begin{pmatrix}\mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n\end{pmatrix} \tag{A3.64}$$

The dimensions of these matrices are

$$(3n \times 7)(7 \times 1) + (3n \times 6n)(6n \times 1) = (3n \times 1)$$

The alternative form of the seven-parameter similarity transformation is the one that achieves a rotation about a local origin, $\mathbf{x}_0$, by reformulating the transformation equations as

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} + \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} + (1+\lambda)\mathbf{R}\begin{pmatrix} x - x_0 \\ y - y_0 \\ z - z_0 \end{pmatrix} \tag{A3.65}$$

The derivation of the least squares formulation of this transformation proceeds in a very similar way to the original form. The functional relationship is altered to

$$\mathbf{F}_i = \Delta\mathbf{X} + \mathbf{x}_0 + (1+\lambda)\mathbf{R}(\mathbf{x} - \mathbf{x}_0) - \mathbf{X} \tag{A3.66}$$

The matrix $\mathbf{C}_i$ remains the same as before, but $\mathbf{A}_i$ becomes

$$\mathbf{A}_i = \begin{pmatrix} x - y_0 & 0 & -z + z_0 & y - y_0 & 1 & 0 & 0 \\ y - y_0 & z - z_0 & 0 & -x + x_0 & 0 & 1 & 0 \\ z - z_0 & -y + y_0 & x - x_0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{A3.67}$$

The vector $\mathbf{b}$ is evaluated with the functional relationship in (A3.66), but all the matrices are combined as in (A3.64).

For either of these two models, the weight matrix is a three-dimensional version of that given in (A3.33) for the two-dimensional case.

A3.3.2 Other similarity transformations

In some situations, a full seven-parameter similarity transformation is either not needed or not possible with the existing data. It may be the case, for example, that only the average shift, or translation, between the datums is required. This would then be termed a three-parameter transformation, with only (ΔX, ΔY, ΔZ) being determined.

To take another example, GPS data could be transformed onto a local datum whilst preserving the scale of the survey by applying a six-parameter transformation solving for (α_1, α_2, α_3, ΔX, ΔY, ΔZ). Each of these cases is simply derived as a special case of the seven-parameter transformation, in which the columns of the matrix **A** that correspond to the unwanted parameter are eliminated. Thus, the design matrix for a solution of ($\Delta X, \Delta Y, \Delta Z$) alone is found by eliminating the first four columns of (A3.61):

$$\mathbf{A}_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{A3.68}$$

Similarly, the design matrix for a six-parameter transformation involving three shifts and three rotations is found by eliminating the first column of (A3.61):

$$\mathbf{A}_i = \begin{pmatrix} 0 & -z & y & 1 & 0 & 0 \\ z & 0 & -x & 0 & 1 & 0 \\ -y & x & 0 & 0 & 0 & 1 \end{pmatrix}$$

In each case, the design matrix **C** will remain the same as in (A3.62).

A3.4 Worked example

The way in which the solution is carried out is illustrated here with a worked example. The one selected is a three-dimensional similarity transformation by the two different models shown in section A3.3.1, although this will also serve to illustrate the general principles for any transformation.

There are four points common to the two systems, and their coordinates are given in Table A1. The standard error of each coordinate in system P is 0.01 m, and in system Q is 0.02. For convenience, it will be assumed that all coordinates and points are uncorrelated, and that the weight matrix is therefore diagonal. It has the form

$$\mathbf{W} = \begin{pmatrix} \mathbf{w} & 0 & 0 & 0 \\ 0 & \mathbf{w} & 0 & 0 \\ 0 & 0 & \mathbf{w} & 0 \\ 0 & 0 & 0 & \mathbf{w} \end{pmatrix}$$

where each sub-matrix **w** represents one of the points and is given by

$$\mathbf{w} = \begin{pmatrix} 1/0.01^2 & & & & & & \\ & 1/0.01^2 & & & & & \\ & & 1/0.01^2 & & & & \\ & & & 1/0.02^2 & & & \\ & & & & 1/0.02^2 & & \\ & & & & & 1/0.02^2 & \end{pmatrix}$$

Table A1

System P			
Point	*X*	*Y*	*Z*
A	4 027 656.73	702.96	4 973 741.92
B	4 025 033.77	14 050.08	4 975 857.89
C	4 010 282.95	1 399.85	4 987 786.36
D	4 009 387.42	13 295.68	4 988 482.31
System Q			
Point	*X*	*Y*	*Z*
A	4 027 756.52	820.90	4 973 972.92
B	4 025 134.97	14 168.85	4 976 087.46
C	4 010 381.77	1 521.26	4 988 016.66
D	4 009 487.60	13 417.55	4 988 711.44

As a first approximation, the provisional values of all the transformation parameters are set to zero, which will always be a valid assumption for any three-dimensional geodetic datums.

Each element of the design matrix **A** is then formed according to (A3.61), using the observed values of the coordinates. The full matrix is:

$$\begin{pmatrix}
4027656.7 & 0.0 & -4973741.9 & 703.0 & 1 & 0 & 0 \\
703.0 & 4973741.9 & 0.0 & -4027656.7 & 0 & 1 & 0 \\
4973741.9 & -703.0 & 4027656.7 & 0.0 & 0 & 0 & 1 \\
4025033.8 & 0.0 & -4975857.9 & 14050.1 & 1 & 0 & 0 \\
14050.1 & 4975857.9 & 0.0 & -4025033.8 & 0 & 1 & 0 \\
4975857.9 & -14050.1 & 4025033.8 & 0.0 & 0 & 0 & 1 \\
4010282.9 & 0.0 & -4987786.4 & 1399.9 & 1 & 0 & 0 \\
1399.9 & 4987786.4 & 0.0 & -4010282.9 & 0 & 1 & 0 \\
4987786.4 & -1399.9 & 4010282.9 & 0.0 & 0 & 0 & 1 \\
4009387.4 & 0.0 & -4988482.3 & 13295.7 & 1 & 0 & 0 \\
13295.7 & 4988482.3 & 0.0 & -4009387.4 & 0 & 1 & 0 \\
4988482.3 & -13295.7 & 4009387.4 & 0.0 & 0 & 0 & 1
\end{pmatrix}$$

The sub-elements of the design matrix **C** are as defined in (A3.62), which with all transformation parameters having provisional values of zero evaluates as

$$\mathbf{C}_i = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

These sub-matrices combine into a full **C** matrix of dimensions 12×24, as shown in (A3.64).

The vector **b** is then evaluated from (A3.63), using the provisional values of the transformation parameters (all zero in the first iteration) and the observed values of

the coordinates. Thus, the first sub-vector, $\mathbf{b}_1$, is given as

$$\begin{pmatrix} 99.79 \\ 117.94 \\ 231.00 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + (1+0) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4027656.73 \\ 702.95 \\ 4973741.92 \end{pmatrix} - \begin{pmatrix} 4027756.52 \\ 820.90 \\ 4973972.92 \end{pmatrix}$$

and the full vector, including a similar computation for each point, is

$$\begin{pmatrix} 99.79 \\ 117.94 \\ 231.00 \\ 101.20 \\ 118.78 \\ 229.57 \\ 98.82 \\ 121.41 \\ 230.31 \\ 100.19 \\ 121.87 \\ 229.13 \end{pmatrix}$$

All necessary matrices have now been formed, and it remains only to carry out the matrix algebra solution of (A3.13) to determine the best estimates of the transformation parameters. This gives the solution shown in Table A2.

Table A2

λ	2.0691^{-5}	= 20.69 ppm
α_1	9.90267^{-5} rad	= 20.42″
α_2	4.99385^{-5} rad	= 10.30″
α_3	0.00011886 rad	= 24.52″
ΔX	264.75	= 264.75 m
ΔY	104.14	= 104.13 m
ΔZ	−73.00	= −73.00 m

The solution so obtained is added to the provisional set of values (in this case zero) to find the best estimates of the transformation parameters. If the two datums were very different from each other, in theory a further iteration would be required in which the current best estimates of the transformation parameters become the provisional values. In most situations, however, this would not be required.

The alternative to this transformation model is that defined in equation (A3.66). Although the full numerical derivation of this will not be given here, it is instructive to examine the final result. This gives the transformation parameters as shown in Table A3.

What is noticeable about this set of parameters is that, although the scale and the rotations are exactly the same as before, the translations are completely different. This is due to the fact that the origin of the rotation in the first model was the centre of the Earth, rather than being in the region of the points being transformed. Effectively this

Table A3

λ	2.0691^{-5}	= 20.69 ppm
α_1	9.90267^{-5}	= 20.42″
α_2	4.99385^{-5}	= 10.30″
α_3	0.00011886	= 24.52″
ΔX	100	= 100 m
ΔY	120	= 120 m
ΔZ	230	= 230 m

means that as the rotation origin is at such a great distance, its effect is very similar to a translation. The true translation is therefore altered to take account of this.

Although it must be emphasised that the *effect* of these two models is exactly the same (they would give the same coordinates if used to carry out a transformation) the second approach usually gives a better insight into the physical differences between the two datums: the translation found does correspond to the origin shift between the two systems. In the first approach the translation is so dependent on the rotation, and the rotation so dependent on small coordinate errors, that very different answers will be found for different groups of points.

It must be stressed that the two sets of transformation parameters are not interchangeable: if a set of parameters is quoted, the model used must be given alongside them.

References

Allan, A.L. (1997a) *Maths for Map Makers.* Whittles Publishing, Latheronwheel, Caithness.

Allan, A.L. (1997b) *Practical Surveying and Computations*, revised 2nd edn. Laxton's, Oxford.

Beazley, P.B. (1994) *Technical Aspects of Maritime Boundary Estimation.* (Maritime Briefings 1(2)). International Boundaries Research Unit, University of Durham.

Bomford, G. (1980) *Geodesy*, 4th edn. Clarendon Press, Oxford.

Boucher, C. and Altamimi, Z. (1998) Specifications for reference frame fixing in the analysis of a EUREF campaign.
ftp://lareg.ensg.ign.fr/pub/euref/info/guidelines/REF.FRAME.SPECIFV4

Calvert, C. (1994) Great Britain – Changing the National Grid and Geodetic Datum. *Surveying World*, **2**(6), 23–24.

Cooper, M.A.R. (1987) *Control Surveys in Civil Engineering.* Collins, London.

Cross, P.A. (1983) Advanced least squares applied to position fixing. Working Paper No. 6, University of East London.

DMA (1997) *DoD World Geodetic System 1984: Its Definition and Relationships with Local Geodetic Systems*, 3rd edn. Defense Mapping Agency, US Department of the Interior, Washington, DC.

Forsberg, R. and Tscherning, C.C. (1981) The use of height data in gravity field approximation by collocation. *Journal of Geophysical Research* **86**(B9), 7843–7854.

GeoTIFF (1995) Format Specifications Home page at
http://home.earthlink.net/~ritter/geotiff/spec/geotiffhome.html

GIAC (1999) GPS Interagency Advisory Council. Home page at
http://www.ngs. noaa.gov/FGCS/GIAC/giac.html

Higgins, M.B., Broadbent, G.J., and Martin, R. (1998) The use of GPS for the maintenance and investigation of tidal datum: a case study in Queensland, Australia. *Proceedings of the XXI Congress of the Fédération International des Géomètres*, Brighton.

Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J. (1997) *GPS: Theory and Practice*, 4th revised edn. Springer-Verlag, New York.

Hooijberg, M. (1997) *Practical Geodesy Using Computers.* Springer-Verlag, New York.

IERS (1998) *Annual Report for 1997.* International Earth Rotation Service, Observatoire de Paris, Paris.

Jones, M.A.B. (1990) *Bear Essentials, or Geodesy Facts You Always Wanted to Know But Didn't Know Where to Find Them.* Available from Malcolm Jones, 89 Woodhall Street, Stirling WA 6021, Australia.

Lambeck, K. (1988) *Geophysical Geodesy.* Clarendon Press, Oxford,

Langley, R.B. (1997) The GPS error budget. *GPS World* **8**(3), 51–56.

Leick, A. (1990) *GPS Satellite Surveying*. John Wiley, New York.

Misra, P.N. and Abbot, R.I. (1994) SGS85–WGS84 transformation. *Manuscripta Geodaetica* **19**, 300–308.

Moritz, H. (1980) *Advanced Physical Geodesy*. Wichmann, Karlsruhe.

NASA (1996) EGM96: Earth Geopotential Model 1996. Home page at http://cddis. gsfc.nasa.gov/926/egm96.html

Nyamangunda, P. (1997) Towards a 10 cm local gravimetric geoid for Zimbabwe using FFT. PhD Thesis, University of London.

Ordnance Survey (1995a) *National Grid/ETRF89 Transformation Parameters*. Geodetic Information Paper No 2. Ordnance Survey, Southampton.

Ordnance Survey (1995b) *The Ellipsoid and the Transverse Mercator Projection*. Geodetic Information Paper No 1. Ordnance Survey, Southampton.

Ordnance Survey (1997) *Positional Accuracy of Large-scale Data and Products*. Consultation Paper 3. Available at http://www.ordsvy.gov.uk/literatu/infopapr/1997/cons0397.html

Ordnance Survey (1998) *Global Positioning Systems and mapping in the twenty-first century*. Information Paper 12. Available at http://www.ordsvy.gov.uk/literatu /infopapr/1998/pap1298.html

Rapp, R. and Pavlis, N. (1990) The development and analysis of geopotential coefficient models to spherical harmonic degree 360. *Journal of Geophysical Research*, **95**(B13), 885–911.

Rapp, R.H. (1994) Separation between reference surfaces of selected vertical datums. *Bulletin Géodesique* **69**, 26–31.

Rapp, R.H. and Nerem, R.S.A. (1994) Joint GSFC/DMA project for improving the model of the Earth's gravitational field. In Suenkel, H. (ed.), *Gravity and Geoid, International Association of Geodesy Symposia*, Vol. 113. Springer-Verlag, New York.

Rossbach, U., Habrich, H., and Zarraoa, N. (1996) Transformation parameters between PZ-90 and WGS 84. *Proceedings of the 9th International Technical Meeting of the Satellite Division of the Institute of Navigation*, Kansas City, Missouri, 17-20 September, pp. 279–285.

Seeber, G. (1993) *Satellite Geodesy*. Walter de Gruyter, Berlin.

Sharma, S.K. (1966) A note on geodesic lengths. *Survey Review* **140**, 291–295.

Snyder, J.P. (1987) *Map Projections: A Working Manual*. Geological Survey Professional Paper 1395, United States Government Printing Office, Washington DC.

Stansell, T.A. (1978) *The Transit Navigation Satellite System*. Magnavox (USA).

Torge, W. (1991) *Geodesy*, 2nd edn. Walter de Gruyter, Berlin.

University of Zurich, (1989) *Map projections for SAR Geocoding*. Technical Note 22910, Remote Sensing Laboratories, University of Zurich.

USGS (1999) PROJ.4: Cartographic projection programs. Available at http://kai.er. usgs.gov/ftp/index.html

Walsh, D. and Daly, P. (1998) Precise positioning using GLONASS. *Proceedings of the XXI Congress of the Fédération International des Géomètres*, Brighton.

Willis, P. *et al.* (1989) Connection of the two levelling system datums IGN69 and ODN through the Channel by using GPS and other techniques. *First International Workshop on Geodesy for the Europe–Africa fixed link feasibility studies in the Strait of Gibraltar*, Madrid, 8–10 March.

Index

absolute coordinates in WGS84 52
absolute positions 38
accuracy 38
affine transformation 100, 133
AHD71 20
Airy 47
Albers (conic) equal area 88
angular distance from the central
    meridian 75
anti-spoofing 34
Arc/Info 70, 74, 99, 101
ArcView 101
astronomical azimuth 26
astronomical coordinates 16
atmospheric refraction 39
Australia 20, 109
Australian Height Datum 1971 20
azimuth 73, 123–4
azimuthal equal area projection 81, 102
azimuthal equidistant projection 79–81
azimuthal projections 79 *et seq.*
azimuthal stereographic projection 62, 82
bench marks 56, 106
    in the GPS survey 56
bias terms 41
boundaries 12, 124
British datums 26
British National Grid 59
Bureau International de l'Heure 25
carrier waves 34, 41
cartesian coordinates 5, 12–13
cartesian 8
case studies 105 *et seq.*
central meridian scale factor 76, 93
central meridian 75, 104
    angular distance from 75
centre of mass of the Earth 23
change in spheroidal height 21
change in the geoid separation 22
changes in geodetic coordinates 27
Clarke 1880 47
Clarke's 'best formulae' 125
coarse acquisition code 34
compatibility 95
computational approximations 30
computational aspects 65
computations 58, 93
    in a spheroidal coordinate system 11
    within the coordinate system5
conformal projection 63, 66
conic equal area projection 102
conic equidistant projection 87–8, 102
conic projection 62, 86 *et seq.*, 103
constant value, geoid–spheroid
    separation 18
constellation of satellites 33
control segment of GPS 33
convergence of the Earth's equipotential
    surfaces 19
convergence 59, 62, 75, 104, 111
conversion from geodetic to Cartesian
    coordinates 6
conversion of data 6–7
coordinates 3–6
    of the geodesic point 127
    of the origin of the local datum 27
    polar 4
    rectangular 4
    required in the local datum 40
correction to be applied to angles 65

covariances 133
crust and mantle 14
curvilinear 8
cylindrical equal area projection 70–2
cylindrical equidistant projection 68–70, 102, 112
cylindrical projections 68 *et seq.*
cylindrical surface 62
datum 3–6
  British 26, 28
  definition 3, 26
  European 26, 29
  geocentric 23
  global 23
  local 26
    distortions in 30
  national 26
  realisation of 30–2
  satellite 23–6
  transformations 45
  transformations for precise applications 50
Defense Mapping Agency 47
defining parameters of projections 58–9
definition of scale factor 60
degree 16
delta errors 36
design matrices 129, 135
design matrix 136
designing a projection 58, 66
determination of transformation parameters by least squares 127
developable surfaces 58, 61
deviation of the vertical 16, 26
differential GPS (DGPS) 38, 39, 40
differential spheroidal height 40
direct conversion 6
direct transformations 95 *et seq.*
distances 123
distortions
  in the local datum 30, 53
  in satellite images 98
drafting errors 98
dual frequency observations 36
Earth Geopotential Model 1996 (EGM96) 17
Earth rotation 100
Earth's crust and mantle 17
eastings 60
eccentricity 11
effects of an error in separation 46
EGM96 41, 47
electromagnetic distance measurer (EDM) 65
ellipsoid 9
  of revolution 9
ellipsoidal height 20, 22, 38
ephemeris 33, 38
ephemeris error 36
epsilon errors 36
equal area projection 63, 66, 102
equidistant projection 63
equipotential surface 14
error in the satellite ephemeris39
error
  delta 36
  drafting 98
  ephemeris 36
  epsilon 36
  in digitising 98
  sources of 36
  survey 98
errors in the satellite clock 36
estimating the transformation parameters 47
ETRF89 26, 32
European datums 26
European Terrestrial Reference Framework 26
false coordinates 58, 70
false eastings 70, 93, 104
false northings 70, 93, 104
features, preserved 63–4
flattening 10, 11, 59
formulae to convert between geodetic and projection coordinates 59
function of the map 64
functional relationship 128
fundamentals of map projections 58
Gauss-Kruger 77
general form of polar projection 79
general conic projections 86
generalisationin map production 98
geocentricdatum 23
geodesic 124, 126, 127
geodesics 124–5
geodetic azimuth 26
geodetic coordinates 5–6, 11–13, 16, 58, 123
  transformation to Cartesian 12

transformation between datums 6
Geodetic Reference System 1980 (GRS80) 23
geodetic surveys 12
geographic coordinates 11
geographic information systems (GIS) 1, 99, 101
geographical area to be mapped 64
'geographical coordinates' 111
geographical extent of region to be mapped 62
geoid 5, 14 *et seq*, 20, 107, 108
and the spheroid 6
modelled as a uniform slope 19
problems 53
geoid–spheroid separation 6, 14, 18, 26
geometric distortions 100
geometry of the ellipsoid 123
GeoTIFF format 59
global Earth model 16
global geodynamic studies 42
global geopotential models 20
global positioning system (GPS) 1, 12, 24, 33 *et seq.*, 93, 123
data 115
two-dimensional tansformation 115
phase measurements 41
Global Navigation Satellite System (GLONASS) 24
gnomonic projection 83–4, 103
graticules 58, 59–60
gravity 14
observations 16
great circle 73, 83, 123
Greenwich Mean Time 9
Greenwich meridian 12
Greenwich 8
grid north 75
grids 58, 59–60
ground control 97–8
ground control points 97
ground survey techniques 30
height above mean sea level 19
height 8
reference surfaces for 19–22
heights above the geoid 5
Hotine oblique Mercator projection 78
hybrid of transformation 109
ideal geoid 20
influence of the geoid in deriving coordinates 51
integer ambiguities 42
integer ambiguity 41
international celestial reference system 25
International Earth Rotation Service (IERS) 25
International GPS for Geodynamics Service 24, 41
International Polar Motion Service 25
International Terrestrial Reference Framework (ITRF) 25, 32
international terrestrial reference system 25
interpolation algorithm 108
ionospheric refraction 36, 43
irregularities of the terrain 17
ITRF *see* International Terrestrial Reference Framework
Jacobian matrices 129
kinematic phase GPS 43
kinematic techniques 43
knowledge of datums transformation parameters 47 *et seq.*
knowledge of height, H 46
knowledge of separation, N 45–6
Lambert conformal conic projection 62, 88–90, 102, 110, 112, 117
Lambert 88
land surveying 65
computations 64
Laplace azimuth observations 26
large scale mapping 64
latitude of the origin 93
latitude, longitude and height 11
latitude 5, 8
least squares 127
linearisation 129
local and regional datums 26
local datum 26
local reference system 18
local site grid 116
longitude of the origin 93
longitude 5, 8
loxodrome 73
Malaysia 78
map digitising 98
map production 98
errors in 98
map projection 8, 58

fundamentals 58 *et seq.*
map scale 61
map shrinkage 100
maritime boundaries 111
Marseilles 20
mean sea level 5, 14, 20
measurement errors 30
Mercator projection 66, 72–4, 102, 117
meridians 9, 59
mid-latitude regions 87
mineral concessions 12
minimising distortion 90
misclosure vector 132, 134, 135
model for the transformation equations 52
Molodensky's formulae 6, 27
monitoring of movement 22
movement of Earth's poles 25
multipath 36, 38, 43
National Aeronautics and Space Administration (NASA) 17
national datums 26
National Imaging and Mapping Authority (NIMA) 17
navigation chart 74
navigation 66, 72
New Zealand map grid 62
Newlyn 20
noise 38
non-uniform change in the geoid 56
normal cylindrical projections 103
normal section 124
normal section azimuths 124, 126
normal sections 124–5
northings 60
oblate spheroid 9
oblique Mercator projection 62, 77–8
oblique projections 93
offshore users 40
Ohio State University 17
'on the fly' (OTF) techniques 44
Ordnance Survey 31, 94
origin for the projection 69
origin of latitude 103
origin of longitude 103, 110
origin of the projection 69, 70
orthometric height 5, 6, 19, 20, 53
orthomorphic projection 63
OS(GPS)93 32
OSGB36 32, 47
OSU91 17
overall scaling 74, 79, 87
overall scaling factor 93
paper distortion 98
parallels of latitude 9
parallels 59
parameters of the projection 92
parameters of the spheroid 47
parameters 92
permanent reference stations 40
Peters projection 72
plane transformations 98–101
plate carrée 69
point of origin 48
offset from centre of Earth 26
point scale factors 61
'polar' axis 12
polar equal area projection 82
polar equidistant projection 102
polar form of coordinates 4
polar projection 79
general form 79
polar projections 103
polar stereographic projection 102
positional dilution of precision (PDOP) 37
precise code 34
preserved features 63–4
principal differences between WGS84 and WGS72 24
PROJ.4 92
projected coordinate system 12, 53, 59
projection
Albers (conic) equal area 88
coordinates 5, 6
defined 58
defining parameters 58–9
general conic 86 *et seq*
Hotline oblique Mercator 78
Lambert conformal conic 88
Mercator 66, 72, 102, 117
method 59
oblique Mercator 62, 77
origin 58
parameters of 92
Peters 72
polar equal area 82
polar equidistant 102
polar stereographic 102
space oblique Mercator 78

spherical or spheroidal models in 58
summary of parameters 59
universal polar stereographic 82
prolate spheroid 9
provisional values in transformation 133
pseudo-ranges 35
PZ90 24, 25
radius 8
ratio of scale factor 96
real time 39
realisation of a datum 30–2
realise the datum 30
realising a reference system 23
real-time kinematic (RTK) 43
receiver clock 35
receiver noise 36
rectangular form of coordinates 4
redundant data 53
reference receiver 39
refraction 36, 38
refractive index 36
relative positions 38
re-scaling 58
residuals 128
reverse computation 12
of azimuth 126
rhumb line 73
rotation about a local origin, $\mathbf{x}_0$ 137
rotation of the Earth 25
rotations between the coordinate systems 40
rotations 6
rover 39
rubber sheeting algorithm 117
satellite
(or airborne) remote sensing 1
altimetry 16
ephemerides 24
geodesy 26
image 93, 95, 98
orbits 16
reference system 23
time system 35
scale change 6–7
scale factor 30, 58, 60, 61, 62, 65, 66
distortion minimised 110
for each line 65
second-order polynomials 100, 134
selective availability 36, 38
semi-major axis 10
semi-minor axis 10
separation, N 6, 14, 16, 45
seven-parameter similarity transformation 106, 135, 137
seven-parameter transformation 138
SGS85 24
SGS90 24, 25
shape and size of the Earth 8
shifts of the geoid 56
similarity transformation 52, 97, 108
simple two-dimensional transformations 48
Simpson's rule 65
six-parameter transformation 138
software 92
sources of error 36
sources of information 47
Soviet Geodetic System 1985 (SGS85) 24
space oblique Mercator projection 78
sphere 8, 9, 59
spherical coordinates 9, 123
spherical harmonic expansion 16
spherical models 58
spheroid 5, 9
spheroidal (or ellipsoidal) heights 5, 45
spheroidal models58
spheroidal normal 11, 16
spin axis of the Earth 30
standard model of the geoid 22
standard parallel 82, 86, 93, 110
states 12
boundaries in between 12
summary of information required 92 *et seq.*
summary of the parameters 59
survey computations 65
survey errors 98
sympathy with existing mapping 50
temperature, pressure and humidity 36
terrestrial reference frame 25
the geoid 14 *et seq.*
the global positioning system 33
the GPS constellation 34
the similarity transformation 130
three-dimensional shift 6
three-dimensional system 8
cartesian 8
curvilinear 8
three-dimensional transformation 30, 135
three-parameter transformation 138

tide gauge 20
time datum 8
transformation 109
  about a local origin 52
  affine 49
  direct 95 *et seq.*
  hybrid 109
  of geodetic coordinates 27
  of GPS data into a local datum 105
  models 51
  plane 98
  polynomial 49
  process 6
transformation parameters 31, 47
  by least squares 7
transverse cylindrical 74
transverse Mercator projection 62, 65, 74–7, 103, 104, 115
triangulation pillars 106
triangulation 30, 66
tropospheric refraction 36
true eastings and northings 70
true north 77
two- and three-dimensional coordinate systems 8 *et seq.*
two standard parallels 87, 90
two-dimensional affine transformation 100
two-dimensional coordinate system 58
two-dimensional similarity transformation 96, 99
two-dimensional transformation of satellite imagery 113
two-dimensional transformation 95, 115, 130
two-dimensional map projection 8
undulation of the geoid 16
uniform separation of the geoid 54
uniform slopes 56
United States Geological Service 92
universal polar stereographic projection (UPS) 82
universal transverse Mercator system (UTM) 59, 76
unknown projections 101, 116
UPS 93
US government policy regarding selective availability 37
location of utilities in GPS 50
UTM 77, 93
vertical datum 5, 6, 21
vertical reference surface 20
W code 34
'wander' of the pole 30
warping 95
weight matrix 129, 137
WGS72 24
WGS84 (*see also* World Geodetic System) 18, 24, 25, 47
worked example, three-dimensional similarity transformation 138
World Geodetic System 1984 (WGS84) 23,
Y code 34
zones 66
zoning 59

datums and map projections

The development of geographic information systems for handling and manipulating data in digital form, and also the development of techniques such as the global positioning system and satellite (or airborne) remote sensing, has led to a vast increase in the use of spatial data. This book is a practical guide for those working with spatially referenced data to the problems that may be associated with datums and map projections. The book makes the issues clear without assuming any prior knowledge and focuses on solving the problems encountered when combining data from different sources. It explores the short cuts applicable when incomplete information is available. There are many practical examples and extensive case studies and appendices of essential formulae. It is ideal for students and practitioners in surveying, remote sensing, geographic information systems and related areas and caters for the non-specialist.

- clear guide – assumes no prior knowledge
- includes GPS and how GPS-derived data can be combined with other sources of spatial data
- many practical examples and case studies
- problem-solving approach

whittles publishing

ISBN 1-870325-28-1

9 781870 325288

CRC

ISBN 0-8493-0884-4

90000

9 780849 308840