COORDINATE SYSTEMS AND MAP PROJECTIONS FOR GIS

D H MALING

The subjects of coordinate systems and map projections are treated under three major headings. First, it is necessary to emphasize the need for economical methods of handling GIS data and to describe some of the ways in which economies may be introduced to the transformation processes. Secondly, there is a short account of some of the methods of transformation which may be used in GIS. Thirdly, there is a description of a method of choosing suitable projections for particular GIS applications.

INTRODUCTION

This chapter is concerned with a review of the principal methods which may be used to transform positional data so that they may be registered with other positional data and so that the results of analyses can be output as maps. The terminology relating to map projections is that used by Maling (1968, 1973), Royal Society (1966) and ICA (1973); the appropriate theoretical background is to be found in Richardus and Adler (1972) and Maling (1973). Because this chapter is wholly concerned with geometric transformations applied to positional data, unqualified use of the word data in this chapter refers to the positions of points on a map, photograph, remotely sensed image, or in a file.

Figure 10.1 illustrates the various types of coordinate system which are used in this chapter. The discussion proceeds from the initial assumption that the primary sources for GIS positional data are printed maps which have been converted by digitizing into machine readable form. This information may be converted into either three-dimensional terrestrial coordinates or two-dimensional plane coordinates. In the first form these are either geographical coordinates of latitude and longitude, (φ,λ) or three-dimensional Cartesian coordinates (X, Y, Z). In the second form the stored data are referred to a plane coordinate system. This may be simple plane Cartesian, polar coordinates, a raster grid or a map projection. At first sight it seems sensible to use terrestrial coordinates as the preferred method of storing data. However, the objection to relying upon this procedure is the sheer volume of data which needs to be handled and stored. A file representing a vector digitized map may comprise many tens of thousands of points. For example, Cocks, Walker and Parvey (1988) have described the contents of the GIS of Australia (AIS) in which each map base comprises 20 000 coordinate pairs for the low resolution outline and 300 000 coordinate pairs for high resolution use. A Landsat Multi-spectral Scanner (MSS) image for only one waveband comprises more than 7 million pixels; a complete Landsat Thematic Mapper (TM) image (seven bands each comprising 5700 lines of 6900 pixels) occupies 262 Mb of storage. From the point of view of handling these data economically in the transformation from geographical position through the formation and registration of layers, it is desirable to transform the raw data extracted from map sources into a uniform system of positional referencing within the system itself. This removes the need for preliminary processing of each layer every time it is registered to another layer. This is especially important if there is a mixture of vector and raster data to be standardized.

One of the commonest solutions is that of the

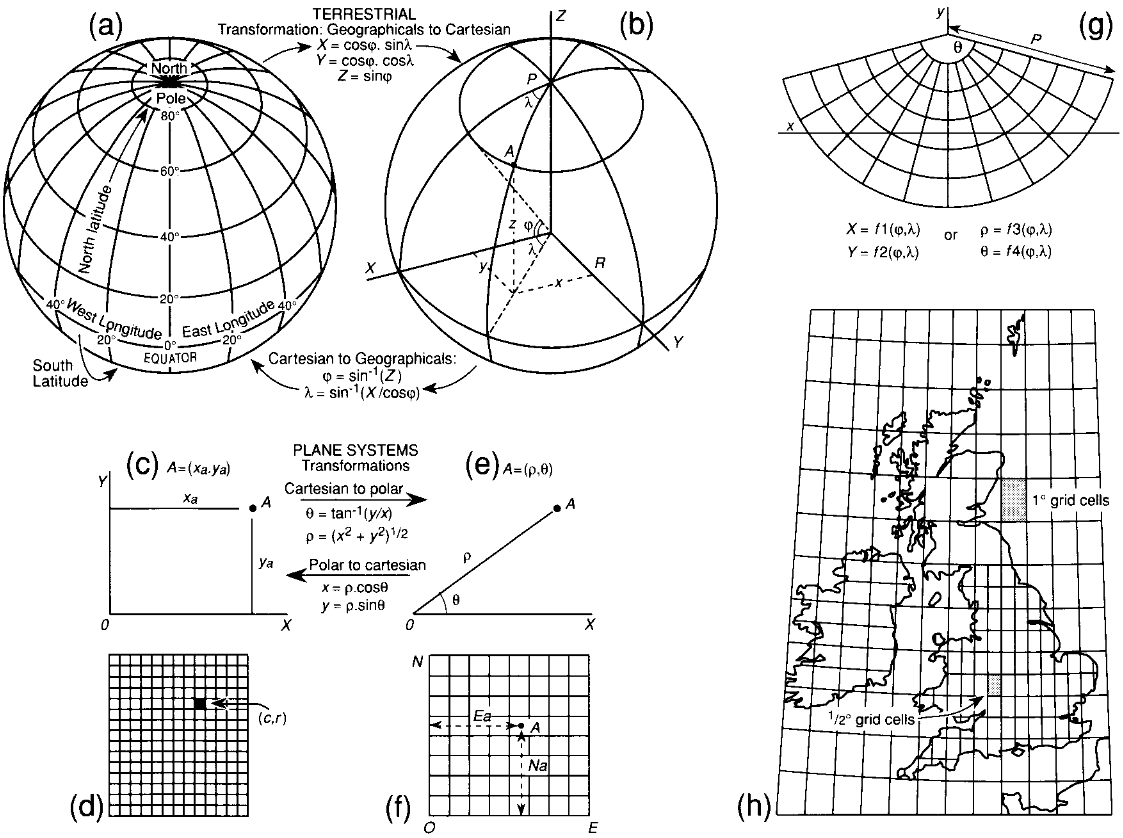

Fig. 10.1 The coordinate systems used in this chapter. (a) and (b) Terrestial systems: (a) Geographical coordinates of latitude (φ) and longitude (λ). (b) Three-dimensional Cartesian coordinates (X, Y, Z). (c) – (h) Plane systems: (c) Plane Cartesian (x, y) coordinates. (d) Raster (c, r) coordinates in which position is determined by counting cells rather than by analogue measurement. (e) Plane polar (ρ, θ) coordinates. (f) A map grid is another example of a plane grid expressed in linear measurement (E, N), which are usually metres on the ground. (g) Graticule in which φ and λ are expressed in either (x, y) or (ρ, θ) coordinates.

At the output stage, map projections are defined either in Cartesian or in polar coordinates vis:

$$x = f_1(\varphi,\lambda),$$
$$y = f_2(\varphi,\lambda),$$
$$\rho = f_3(\varphi,\lambda),$$
$$\theta = f_4(\varphi,\lambda),$$

(h) Grid cells are a compromise between geographical coordinates and a grid. They are created by subdividing the graticule into quadrangles of suitable dimensions. In this example, quadrangles of size 1° × 1° and ½° × ½° are illustrated.

grid cell, comprising a fairly close pattern of spherical quadrilateral cells which are derived from subdivision of the graticule into one-degree or half-degree units. They are, therefore, much larger than a typical raster cell. The Australian AIS makes use of both of these dimensions, together with even larger units (Cocks, Walker and Parvey 1988; see also O'Callaghan and Garner 1991 in this volume). Grid cells are neither square nor rectangular because their sides are formed from two meridians and two parallels. The convergence of the meridians towards the poles means that the grid cells have a pair of sides of different length (Fig. 10.1h). Since great use is made of this unit by the United States Geological Survey (USGS) as the basis for the sheet lines of its maps, the word 'quadrangle' is frequently used to describe it.

GIS FRAMEWORK

The next development from using grid cells to hold data is to use a map projection at this intermediate stage. The concept is so important that it is here called the 'GIS framework' to distinguish it from the methods already considered and any other projection which may be used for purposes of illustration of the same data. This device is obviously of increasing importance the greater the area to be covered by a GIS, for two reasons. First, the database for the whole of a large country or a continent will be large and storage of data in terrestrial coordinates is impractical. Secondly, the area covered will be far too large for any convenient approximations that the earth is flat so a plane grid will not suffice. The GIS framework is most likely to be that used for national surveys, usually the Universal Transverse Mercator (UTM), the Soviet equivalent to UTM, or the Lambert Conformal Conical projection. However, in many instances the coverage of the GIS extends beyond the single state so that the different origins and projections used by different states must be reconciled. Also, the projections used for some thematic maps to be incorporated into the GIS may differ from those of the topographical base. Maps of maritime distributions are almost invariably on different projections to those used for land maps.

Briggs and Mounsey (1989) and Mounsey (1991 in this volume) have emphasized the difficulties which arise in handling numerous and disparate sources in creating CORINE, the environmental database for the European Community. They have cogently argued the need for a common projection framework to relate all the component layers. For a continental GIS it may even be necessary to use more than one type of projection as the framework. In the design of the environmental GIS for Antarctica, Sievers and Bennat (1989) describe the design of a set of Lambert conformal conical and stereographic projections created as raster grids to serve as the mathematical framework for the system.

Economy of data handling

A fundamental principle of conventional cartography is that there is a limit to the smallest size of object which can be shown legibly on a map. This is usually taken to be a map distance of about 0.15 mm and it is often called the zero dimension. It has the important effect of placing a limit upon the degree of complexity which is needed in the design and production of a map (see Fisher 1991 in this volume). If a particular computation or cartographic technique affects the plotted position of a point by less than this amount, a simpler procedure may suffice. Thus the zero dimension permits a series of assumptions to be made about the way in which original surveys are computed and plotted and how much credence should be put on measurements (including digitizing) made from maps. It is also important in making a choice about the suitability of a certain projection for a particular job, even if this is often used in the negative sense of deciding whether the projection of the existing map, though incorrect, will suffice. Since maps are the primary source of data for a GIS, the limit created by the zero dimension and any other imperfections (see Maling 1989) revealed during measurement from it, are transferred to the GIS, irrespective of the degree of sophistication of the methods of data capture used to digitize the maps. This is just reiteration of the fundamental truth that no data are better than their sources.

Assumptions about the shape and size of the earth

Short cuts may be made to computations involving the shape and size of the earth by assuming that its shape is geometrically simpler than it actually is. The first of these assumptions made in surveying and mapping, as described in varying detail in books on geodesy, surveying and map projections (e.g. Richardus and Adler 1972; Maling 1973; Jackson 1980), is that the rather complicated surface of the geoid may be replaced by a reference figure or spheroid. There is considerable temptation to write programs which apply transformations with geodetic precision so that distances and directions between points on the curved surface and plane coordinates for the principal projections used for topographical and cadastral mapping are all referred to a particular spheroid. Although such practices are appropriate to field surveys and simulated maps, they do not necessarily apply to handling those GIS layers whose sources were paper maps. It follows that the spherical assumption is still justified in transforming most map data. For example, Snyder (1985, 1987b), Shmutter (1981) and Doytsher and Shmutter (1981) have all presented formulae for transforming data to and from various projections, all of which have been derived for a spherical earth. Thus the spherical assumption is still as valid as it was in the days before computers made it so easy to refer all calculations to the spheroid.

The simplest assumption of all is, of course, that the earth is flat, so that a satisfactory map can be made by plane surveying. For many local surveys carried out for municipal, civil and mining engineering purposes, the extent of the survey and, therefore, the influence of earth curvature is so small that the plane assumption will suffice. Vincenty (1989) has reconsidered this subject in the light of modern survey practice. Extending the flat-earth argument to photogrammetry, an analogue plotter has a plane datum surface which is simulated by its base carriage. Therefore, the pair of aerial photographs placed within the plotter are referred to a plane datum.

Economy in the design of formulae

The actual formulae used in the transformations may be redesigned for more economical processing. Those which have been taken straight from the literature of geodesy and map projections were originally designed for ease of computation using tables and logarithms. Vincenty (1971), Williams (1982), Snyder (1985) and King (1988) have all demonstrated how the well-known geodetic and projection formulae may be improved for digital processing by a little reorganization. An example of such an improvement is given in eqn [10.24].

An apparent complication of the geodetic and projection formulae used before digital computing is the frequent appearance of the term sin 1″, used to convert from an angle expressed in radians into seconds of arc or vice versa. For an explanation of this computing trick see Maling (1973). The conversion was necessary in the days when tables of trigonometric functions were used and the argument was in degree measure. With digital computers came the subroutines for calculating trigonometric functions which had to be accessed using the angles expressed in radians as the argument. Consequently, the need for making the majority of the conversions disappeared. Nevertheless, many early programs written to compute Transverse Mercator coordinates of the spheroid were copied straight from the literature of the pre-digital era, including all the sin 1″ terms, so that CPU time was wasted in making unnecessary conversions of angles from seconds of arc into radians and back again for no reason. Some of these economies which have so far been described appear to offer a negligible saving when used to transform only a few points, but the cumulative effect of applying them to each point in turn can result in considerable savings in both storage space and CPU time when a whole map is transformed. A number of other economies which have greater impact upon the design of GIS are discussed below.

Economies in map use

Most national topographical maps are based upon conformal projections and, consequently, this has become the commonest base for the GIS framework. However, an equal-area projection would theoretically be a more suitable base for many distribution maps. Therefore, it is appropriate to question whether the difference matters.

In using topographical maps as a source for distribution mapping in a country the size of Britain, the influence of the projection can usually be ignored without any serious consequences. Ordnance Survey maps of Britain are based upon a particular version of the Transverse Mercator projection so that there is a certain amount of area distortion on these maps. However, for maps of mainland Britain the area scale nowhere exceeds the range 0.999 08–1.000 92; in other words it varies from the constant area scale of an equal area projection by less than 0.1 per cent. Since this is likely to be smaller than the errors which arise from the imperfections of the source map, it follows that judgements about density of distribution or measurement of area occupied by different categories of land use, for example, are unaffected by the fact that the map projection used is theoretically incorrect.

Economies through independence from artificial boundaries

Nearly all spatial data collected for administrative and cadastral purposes are recorded in parcels (or polygons) which are parts of the earth enclosed by political, administrative or property boundaries (see Dale 1991 in this volume). This is how they are entered in data files, simply because there is no other way of handling the data initially. However, the polygons thus defined do not usually correspond to other kinds of boundary, such as those of geology, vegetation or land use. Moreover, the artificial boundaries of administrative units are often frequently changed so that much time may be spent in revising and updating these files. Cocks, Walker and Parvey (1988) argue that the difficulties of revising such files are an important objection to using them and that this is sufficient reason for converting, wherever possible, to holding the data in grid cells.

Economies through interpolation algorithms

Alternative methods may be used to eliminate particularly slow computations by introducing interpolation methods. Thus the detail shown in a small square or quadrangle on a map may be transformed to another projection by carrying out the full transformation for the corner points of the figure only and then using interpolation formulae to change the internal detail. This is the digital equivalent of using proportional dividers; it has been particularly well exploited in mapping from remotely sensed imagery. Since the algorithms are described elsewhere (e.g. see Mather 1987), they need not be described here.

THE TRANSFORMATION METHODS

The Cartesian coordinates (x, y) of a point on a map are functionally related to position on the earth's surface expressed in geographical coordinates (φ,λ)

$$\left.\begin{aligned} x &= f_1(\varphi,\lambda) \\ y &= f_2(\varphi,\lambda) \end{aligned}\right\} \qquad [10.1]$$

There are three basic methods of relating (x, y) to (φ,λ) or various forms of plane coordinates used on other maps, aerial photographs or scanned imagery. These are referred to here as:

- Analytical transformation;
- Direct or grid-on-grid transformation;
- Polynomial transformation.

Analytical transformation

Analytical transformation is the most obvious and straightforward solution to the problem of relating Cartesian coordinates on a map to geographical coordinates on the earth's surface. This is because it approximates to the methods of classical cartography, that is, locating and plotting points from their geographical coordinates. In the automated applications, the objective is to convert the (x',y') coordinates of points digitized on a source map into their geographical coordinates. These, in turn, are used to determine the (x, y) coordinates for the GIS framework or to create a new map.

The conversion from geographical coordinates into plane coordinates is the normal practice of constructing a map projection and is regarded as the *forward solution*. The preliminary conversion required to find the geographical coordinates from the (x',y') system of digitized coordinates is correspondingly called the *inverse solution*. Thus the transformation model is:

$$(x',y') \rightarrow (\varphi,\lambda) \rightarrow (x,y) \qquad [10.2]$$

<Inverse solution><Forward solution>

Most of the standard works on map projections only provide the equations for the forward solution. This is because in the days before digital mapping became a practical possibility, only the forward equations were needed to construct a graticule; all subsequent transfer of detail was manual. It was only in the field of topographic mapping, using the Transverse Mercator and Lambert Conformal Conical projections in particular, that the two conversions 'geographicals to grid' and 'grid to geographicals' were likely to be employed and both were provided for by the projection tables. The only comprehensive source for both the forward and inverse equations for the commonly used map projections is Snyder (1987a). This manual also includes worked examples of both computations for spherical and spheroidal assumptions.

The analytical transformation equations for Mercator's projection

The relationship between the forward and inverse coordinate expressions may be exemplified by the sets of equations used to define the normal aspect of Mercator's projection which is the basis of virtually all nautical charts. For the projection of the sphere, eqn [10.3] for the forward solution is to be found in most of the standard works on map projections:

$$x = R.\lambda$$

$$y = R.\mathrm{In}\tan(\pi/4 + \varphi/2) \qquad [10.3]$$

where ln is the natural logarithm (to base ϵ), the longitude, λ, is expressed in radians and the radius of the earth, R, is expressed in millimetres at the scale of the proposed map. In order to express (φ,λ) in terms of (x, y), which is the inverse solution, it is necessary to write

$$\begin{aligned}\varphi &= \pi/2 - 2\tan^{-1}(\epsilon^{-y/R})\\ \lambda &= x/R + \lambda_0\end{aligned} \qquad [10.4]$$

where λ_0 is the datum meridian from which longitudes are measured. Here ϵ is the base of natural logarithms (= 2.718 281 8...). It is written as the Greek epsilon to avoid confusion with the eccentricity of the spheroid, e, in the next three equations.

The first complication which needs to be considered is the corresponding relationships for the projection of the spheroid, having semi-axes a and b with eccentricity e derived from

$$e^2 = (a^2 - b^2)/a^2 \approx 0.0067... \qquad [10.5]$$

For the forward solution of Mercator's projection of the spheroid, eqn [10.3] has to be modified to the corresponding equations

$$\begin{aligned}x &= a.\lambda\\ y &= a.\ln\tan(\pi/4 + \varphi/2)[(1 - e.\sin\varphi)/(1 + e.\sin\varphi)]^{e/2}\end{aligned} \qquad [10.6]$$

For the inverse calculation the equation to find latitude is transcendental, needing an iterative solution

$$\varphi = \pi/2 - 2\tan^{-1}\{t[(1 - e.\sin\varphi)/(1 + e.\sin\varphi)]^{e/2}\} \qquad [10.7]$$

and $t = \epsilon^{-y/a}$ The first trial solution is to find

$$\varphi = \pi/2 - 2\tan^{-1}(t) \qquad [10.8]$$

The result is inserted as φ in the right hand side of eqn [10.7] to calculate a new value for φ on the left-hand side. The process is repeated until the results have converged to a difference between the two values for φ which the user considers to be insignificant and the final value for φ may be accepted. Longitude, is obtained from a simple modification for the λ expression in eqn [10.6], namely

$$\lambda = x/a + \lambda_0 \qquad [10.9]$$

Further transformations

The number of stages in the inverse solution may have to be extended for various other reasons. Because most digitizing is done in Cartesian coordinates and it is sometimes appropriate to deal with a map projection which is best derived in polar coordinates, it may be necessary to change plane rectangular coordinates (x',y') into plane polar coordinates (ρ,δ) before determining the geographical coordinates. Thus the transformation model contains an additional stage, as follows:

$$(x',y') \rightarrow (\rho,\delta) \rightarrow (\varphi,\lambda) \rightarrow (x,y) \qquad [10.10]$$

< Inverse solution ><Forward solution>

Similarly a change in aspect, for example to a transverse or oblique projection, involves yet another stage in the succession of transformations. Change in aspect is commonly effected through the system of (z,α) bearing and distance coordinates (Maling 1973), using spherical trigonometry to convert from (φ,λ) into (z,α). Thus:

$$(x',y') \rightarrow (\rho,\delta) \rightarrow (\varphi,\lambda) \rightarrow (z,\alpha) \rightarrow (x,y)$$

< Inverse solution ><Change in aspect><Forward solution>

[10.11]

An alternative to this method of changing aspect is to use a three-dimensional Cartesian system (X, Y, Z) instead of geographical coordinates to relate positions on the spherical surface. Following the work of Wray (1974), Barton (1976) and Arthur (1978) the change in aspect may also be obtained by rotating these axes through the three Eulerian angles at the centre of the sphere. This time the transformations are:

$$(x',y') \rightarrow (\varphi,\lambda) \rightarrow (X,Y,Z) \rightarrow (X',Y',Z') \rightarrow (\varphi',\lambda') \rightarrow (x,y)$$

< Inverse solution >< Change in aspect > <Forward> solution

[10.12]

where (X',Y',Z') are the rotated coordinates of the point (X, Y, Z).

The advantages and disadvantages of the analytical method

The analytical method is rigorous and it is independent of the size of the area to be mapped. However, it can be inconveniently slow. It seems at first sight that this is no longer a problem; that modern high-speed computers have reduced these considerations to insignificance. However, the clumsiness of the analytical method becomes apparent when applied to large data files. This is well demonstrated by eqn [10.11] relating to change in aspect, where each additional transformation stage may involve either the solution of a separate spherical triangle or, in eqn [10.12], the determination of three-dimensional coordinates and rotation of them for every point on the map.

A further problem, highlighted by Snyder (1985, 1987c), is that the naming of projections on existing maps leaves much to be desired and that even when the correct name has been used, important information such as the positions of the standard parallels in a conical projection or the central meridian of the particular version of the Transverse Mercator projection in use has not been stated. Snyder (1985) has written a program which attempts recognition of the projection in use, based upon the digitized coordinates of nine points (on three parallels and three meridians) of the map, but even this can only distinguish between fairly simple examples.

Direct transformation by the Grid-on-Grid Method

This method does not require inverse solution of the geographical coordinates (φ,λ) of the original map, but is based upon the relation between the rectangular coordinates of the same points on the two projections. This technique was used in traditional cartography for such purposes as regridding or plotting a second grid on a military topographical map – hence the name 'grid-on-grid'. This method is also important in mapping from remote sensing and is the method adopted in modern analytical plotters for use with conventional aerial photography. Practically all the methods of applying geometrical corrections to remote sensing imagery, including that derived from Landsat MSS, Landsat TM and SPOT sensors, utilize such techniques employing ground control points of known position to determine the transformation parameters.

The simplest transformation model is, of course:

$$(x',y') \rightarrow (x,y) \qquad [10.13]$$

Two relatively simple numerical procedures which are commonly employed in the mapping sciences are the linear transformations from one plane Cartesian coordinate system into another. There are two major kinds of transformation: the *linear conformal, similarity* or *Helmert* transformation; and the *affine* transformation. The former is expressed in the general form:

$$x = A + Cx' + Dy'$$
$$y = B - Dx' + Cy' \qquad [10.14]$$

The affine transformation is as follows:

$$x = A + Cx' + Dy'$$
$$y = B - Ex' + Fy' \qquad [10.15]$$

In these equations, the known (or digitized) (x', y') coordinates of a point in one system are transformed into the (x, y) coordinates of the second system, through the use of four or six coefficients A–F. In the Helmert transformation the C and D coefficients are common to both the equations for x and y, but in affine transformation it is necessary to introduce separate corrections for each direction. Both transformations may be resolved into three components:

- Translation of the axes or change of origin, corresponding to the coefficients A and B in both eqns [10.14] and [10.15].
- A change in scale from one grid system to the other.
- The rotation of the axes of one grid system with respect to their directions in the other. These are illustrated in Fig. 10.2.

Helmert transformation

For the Helmert transformation the effects of all three displacements are combined to produce the pair of equations

$$x = (m{\cdot}x'{\cdot}\cos\alpha + m{\cdot}y'{\cdot}\sin\alpha) + A \qquad [10.16]$$

$$y = (-\,m{\cdot}x'{\cdot}\sin\alpha + m{\cdot}y'{\cdot}\cos\alpha) + B \qquad [10.17]$$

where A and B are the coefficients in eqn [10.14] which correspond to the shift in the origin of the coordinates parallel with the x and y axes, the angle α is the rotation of the axes required to make these axes parallel and m is a scale factor. Thus if two points, j and k in the first system correspond to J and K in the second, the ratio of the distances jk/JK must be applied to the first system to bring it to the same scale as the second.

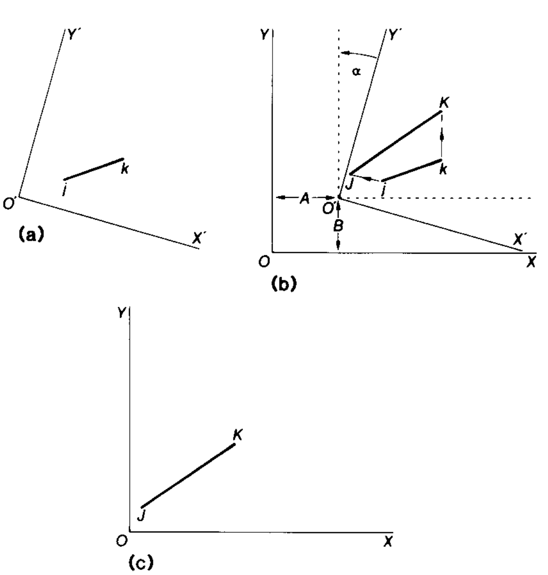

Fig. 10.2 The geometry of Helmert transformation. (a) Initial conditions: showing two points *j* and *k* referred to Cartesian axes $O'X'$ and $O'Y'$ which are orthogonal. (b) The three stages in transformation superimposed upon one another. These are, first the scale change by which the line *jk* is transformed into the line *JK*. Secondly, the rotation of the X' and Y' axes through the angle α about the point O' to make the axes parallel to the final *OX* and *OY* system. Third is the translation of the origin O' through the distances *A* and *B* respectively to refer *J* and *K* to the (X, Y) system. (c) Final conditions: indicating the positions of *J* and *K* within the (X, Y) system.

The complete transformation may be expressed in matrix form as

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} C & D \\ -D & C \end{pmatrix} \cdot \begin{pmatrix} x' \\ y' \end{pmatrix} + \begin{pmatrix} A \\ B \end{pmatrix} \qquad [10.18]$$

where $D = m' \sin \alpha$ and $C = m' \cos \alpha$. The inverse transformation is that of determining the (x', y') coordinates of points whose (x, y) coordinates are already known. Thus

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} C' & -D' \\ D' & C' \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} A \\ B \end{pmatrix} \qquad [10.19]$$

where $C' = \cos \alpha / m$ and $D' = \sin \alpha / m$

If there are only two points (x'_1, y'_1) and (x'_2, y'_2) on the first surface corresponding to (x_1, y_1) and (x_2, y_2) on the second surface whose coordinates are known or have been measured, the method of finding *C* and *D* is through eqns [10.20] and [10.21].

$$C = [\delta x \cdot \delta y' - \delta y \cdot \delta x'] / [\delta x'^2 + \delta y'^2] \qquad [10.20]$$

$$D = [\delta y \cdot \delta y' + \delta x \cdot \delta x'] / [\delta x'^2 + \delta y'^2] \qquad [10.21]$$

where

$\delta x = (x_1 - x_2)$, $\delta x' = (x'_1 - x'_2)$, $\delta y = (y_1 - y_2)$

and

$\delta y' = (y'_1 - y'_2)$

If there are more than two common points, such as occurs in vector digitizing, the adjustment of aerial triangulation or fitting a remotely sensed image to many ground control points, the determination of the coefficients from only two or three of them is inadequate because the coordinates of any of those points may contain small errors, which, in turn, introduces errors into the transformation of all other points. Therefore, all of the data available for the determination of *C* and *D* ought to be taken into consideration. This involves a solution of the coefficients by the method of least squares which is described under the determination of polynomial coefficients.

Affine transformation

The assumption which is made in the Helmert transformation is that the scalar, *m*, is a single unique value. In other words the ratio *jk*/*JK* is the same whatever the directions of these lines. This is a reasonable assumption for some purposes but it may not always be justifiable. For example, in photogrammetry the location of image points on a film may be affected by deformation of the film base by stretching and shrinking and this is not usually the same in all directions. In the extraction of positional information by digitizing a paper map, the influence of differential stretching or shrinking of the paper is even more erratic. For these applications it is desirable to use the affine transformation or even a higher order polynomial because this allows for different scales in the directions of the two axes, m_x and m_y. This may also be combined with small departures of the coordinate axes from the perpendicular, as illustrated in Fig. 10.3. Here it can be seen that the (x, y) axes intersect at an angle $\gamma \neq 90°$. The solution is described in greater detail by Mikhail (1976) and Sprinsky (1987).

Numerical transformation methods

The third method of relating Cartesian coordinates on a map to geographical coordinates on the earth's

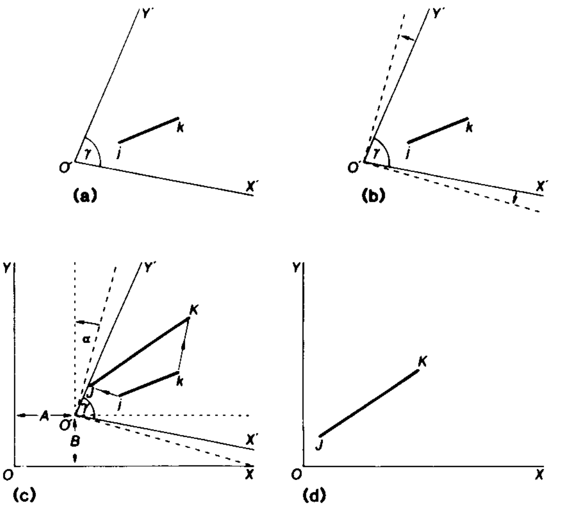

Fig. 10.3 The geometry of affine transformation. (a) Initial condition: showing two points *j* and *k* referred to Cartesian axes *O'X'* and *O'Y'* which are not orthogonal but intersect at some angle γ. The same effect is produced by differential scales in the directions of the two axes. (b) Creation of the orthogonal axes. here shown by the broken line. This is also equivalent to making the linear scale in the *X'* direction equal to that in the *Y'* direction (c) All transformation stages are superimposed. The third, fourth and fifth stages comprise those in the Helmert transformation, i.e. the uniform scale change to represent *jk* by *JK*; the rotation of the axes through the angle α and, finally, the translation through *A* and *B* to refer the points *J* and *K* to the *OX* and *OY* axes. (d) Final transformation illustrates this stage, where *J* and *K* are shown within the *OXY* system.

surface is to construct polynomial expressions to fit the data and use the resulting coefficients to transform the coordinates of the remaining points of map detail. This method is, of course, important in numerical analysis and has many different applications. In the narrower field of transforming positional data for GIS applications this method may be used with equal efficiency for transformation from geographical into grid coordinates (eqn [10.22]) as for making the grid-to-grid transformations (eqn [10.23]).The required number of common points needed to determine the coefficients and the amount of computation needed vary according to the order or degree of the polynomial. For example, a third-order polynomial, containing terms in φ and λ up to φ^3 and λ^3 requires ten coefficients denoted a_{ij}, and the ten in b_{ij} as in eqn [10.22] determined by solving ten or more equations.

The third-order polynomial expression relating grid to geographical coordinates may be written in the form:

$$\begin{aligned} x &= a_{00} + a_{10}\lambda a_{01}\varphi + a_{20}\lambda^2 + a_{11}\lambda\varphi + a_{02}\varphi^2 + a_{30}\lambda^3 \\ &\quad + a_{21}\lambda^2\varphi + a_{12}\lambda\varphi^2 + a_{03}\varphi^3 \\ y &= b_{00} + b_{10}\lambda + b_{01}\varphi + b_{20}\lambda^2 + b_{11}\lambda\varphi + b_{02}\varphi^2 + \\ &\quad b_{30}\lambda^3 + b_{21}\lambda^2\varphi + b_{12}\lambda\varphi^2 + b_{03}\varphi^3 \end{aligned} \qquad [10.22]$$

Similarly the polynomial equations used to transform from grid to grid are:

$$\begin{aligned} x &= c_{00} + c_{10}x' + c_{01}y' + c_{20}x'^2 + c_{11}x'y' + c_{02}y'^2 + \\ &\quad c_{30}x'^3 + c_{21}x'^2y' + c_{12}x'y'^2 + c_{03}y'^3 \\ y &= d_{00} + d_{10}x' + d_{01}y' + d_{20}x'^2 + d_{11}x'y' + d_{02}y'^2 \\ &\quad + d_{30}x'^3 + d_{21}x'^2y' + d_{12}x'y'^2 + d_{03}y'^3 \end{aligned} \qquad [10.23]$$

In pre-computer days polynomial expressions were usually left in this form because it was generally easier to compute each term individually. However, in view of what has already been said about economy in the design of equations, a nested form of each equation may be obtained from a little algebraic rearrangement. For example, the expression for *x* in eqn [10.22] may also be written (Snyder 1985) as:

$$\begin{aligned} x &= a_{00} + \varphi(a_{01} + a_{02}\varphi) + \lambda(a_{10} + \varphi(a_{11} + a_{12}\varphi)) + \\ &\quad \lambda^2(a_{20} + a_{21}\varphi + a_{30}\lambda)\ldots \end{aligned} \qquad [10.24]$$

This example is particularly instructive. Snyder (1985) has reported that the savings which result from using eqn [10.24] rather than the expression for *x* in eqn [10.22] are between 20 and 30 per cent in the solution of a fifth-order polynomial.

Snyder has also shown that conformal projections of the spheroid may be transformed more accurately (therefore requiring a lower-order polynomial) by using the isometric latitude ψ, in place of geodetic latitude in eqns [10.22]. For an explanation of the purpose and use of isometric, and other auxiliary latitudes, the reader is referred to Snyder (1987a) and Richardus and Adler (1972).

Determination of the polynomial coefficients

In order to find the 20 coefficients a_{ij}, b_{ij} in the third-order polynomials above, it is necessary to know the plane rectangular coordinates of 10 corresponding points x_i, y_i and φ_i, λ_i to form the linear equations from which the coefficients can be solved. The amount of data needed to determine the coefficients of a polynomial depends upon the order of the polynomial, which, in turn, depends upon the highest powers of the independent variables used in the terms. For example, first, second, third, fourth and fifth degree polynomials require a minimum of 3, 6, 10, 15 and 21 corresponding points respectively. The common solution is to use even more than these minimum numbers, to obtain the required coefficients by the method of least squares. This is the condition that the sum of squares of differences between the measured and the theoretical coordinates in the new projection should be minimized. Modern textbooks on survey adjustments and computations, for example, Cooper (1974), Mikhail (1976) and Methley (1986) all deal with this subject. The following ($m \times n$) matrix solution is applicable for any number of coefficients, n, and common points, m, but a practical limit is usually created by the capacity of the computer. It is well known in numerical analysis that although a polynomial may be extended to include higher powered terms in $\varphi^4, \lambda^4, \varphi^5, \lambda^5$, etc., the labour of determining the coefficients will hardly justify the extra computing time. Snyder (1985) provides the example of the solution of eqns [10.22] which shows that increasing the degree of the polynomial from third order to fourth order barely justifies the greater accuracy obtained for any purpose other than geodetic work.

In eqns [10.25] and [10.26], the individual coefficients form the column matrix on the left hand side and the control, or common point coordinates are the column matrix on the right hand side.

$$\begin{pmatrix} a_{00} \\ a_{01} \\ \cdots\cdots \\ \cdots\cdots \\ a_m \end{pmatrix} = \mathbf{D} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \cdots\cdots \\ \cdots\cdots \\ x_m \end{pmatrix} \qquad [10.25]$$

$$\begin{pmatrix} b_{00} \\ b_{01} \\ \cdots\cdots \\ \cdots\cdots \\ b_m \end{pmatrix} = \mathbf{D} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \cdots\cdots \\ \cdots\cdots \\ y_m \end{pmatrix} \qquad [10.26]$$

The matrix **D** is calculated from

$$\mathbf{D} = [\mathbf{A}^T.\mathbf{A}]^{-1}.\mathbf{A}^T \qquad [10.27]$$

where the ($m \times n$) matrix **A** is formed from the geographical (or grid) coordinates of the corresponding points. Thus for the third degree polynomial requiring ten terms per line, $n = 10$

$$\mathbf{A} = \begin{pmatrix} 1 & \lambda_1 & \varphi_1 & \lambda^2{}_1 & \lambda_1\varphi_1 & \varphi_1{}^2 & \lambda_1{}^3 & \lambda_1{}^2\varphi_1 & \lambda_1\varphi_1{}^2 & \varphi_1{}^3 \\ 1 & \lambda_2 & \varphi_2 & \lambda^2{}_2 & \lambda_2\varphi_2 & \varphi_2{}^2 & \lambda_2{}^3 & \lambda_2{}^2\varphi_2 & \lambda_2\varphi_2{}^2 & \varphi_2{}^3 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & \lambda_m & \varphi_m & \lambda^2{}_m & \lambda_m\varphi_m & \varphi_m{}^2 & \lambda_m{}^3 & \lambda_m{}^2\varphi_m & \lambda_m\varphi_m{}^2 & \varphi_m{}^3 \end{pmatrix} \qquad [10.28]$$

This solution is due to Wu and Yang (1981) with a fuller derivation by Snyder (1985). The method depends for its accuracy upon the size of the area mapped. This is because a polynomial transformation works well enough with homogeneous data, but a file comprising data digitized from a paper map may not be homogeneous because different parts of it have been affected differently by paper deformation. Just as it is necessary to treat separately the panels of a map which has at some time been folded, it may be necessary to divide the whole map into blocks and transform each block separately.

SOME FACTORS INFLUENCING THE CHOICE OF A SUITABLE PROJECTION

The principles and methods of transformation which have been described are applicable to maps of any scale. However, application of a GIS to a large country or even a continent necessitates choice of a projection, first to serve as the GIS framework and possibly as a suitable projection for displaying the results. It is a fundamental principle of distortion theory that the particular scales and, therefore, exaggeration of areas and angles increase from the origin of the projection towards its edges. Therefore, it is desirable to choose a projection in which either the average or the extreme distortions are small. The amount of distortion on a map depends upon the location, size and shape of the area to be mapped. Distortion is least in the representation of a small, compact country and greatest in maps of the whole world. The three variables – location, size and shape – usually determine the choice of origin, aspect and class of a suitable projection. These may be chosen by the graphical and analytical methods described by Maling (1973). These are based upon the principle that the distortion pattern, its fundamental property, remains constant within a particular projection even when the aspect of the projection is changed. Therefore, the plotted pattern of distortion isograms may be regarded as a frame which can be used to imagine how the distortion will occur, just as an artist may compose a picture by looking at objects through a small rectangular cardboard frame or a photographer uses the rectangular ground glass screen of the camera viewfinder.

In the pre-computer period when the methods were evolving, this was carried out using transparent overlays which were placed singly or in groups over a rough outline sketch map of the country or continent drawn at the same scale. By shifting the position and orientation of the overlay it is possible to estimate any advantage to be gained from a change in origin or change in orientation of the lines of zero distortion. The actual choice of projection depends upon comparison of the patterns of distortion isograms for different projections. When two or more overlays for different projections are superimposed, the extreme values for area scale or maximum angular deformation may be estimated from the isograms. Fig. 10.4 illustrates such a comparison.

It must be realized that the outlines shown on the underlying map are only a rough guide, for the detailed relationship between these and the isograms is only true for that aspect and projection upon which the map was compiled. The purpose of the outline is to indicate approximately the extent of the country or continent; it is the comparison between the distortion isograms which is important.

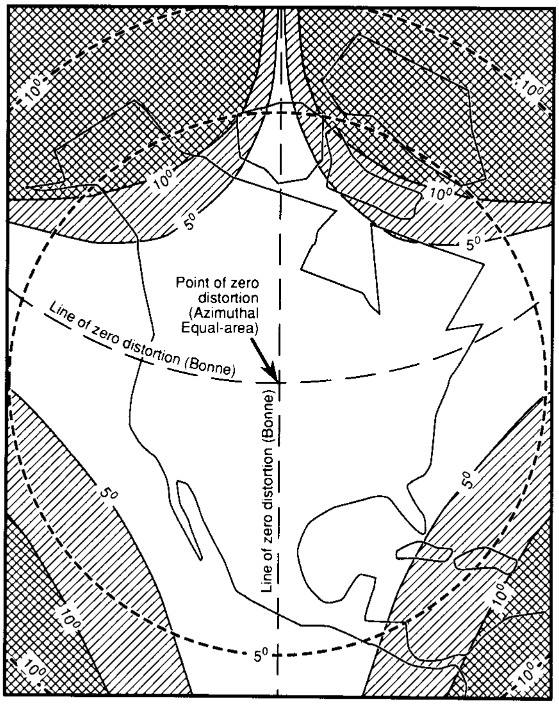
10°
10°
10°
10°
5°
5°
Point of zero distortion (Azimuthal Equal-area)
Line of zero distortion (Bonne)
Line of zero distortion (Bonne)
5°
5°
10°
10°
10°
5°
10°

Fig. 10.4 The comparison of the relative merits of Bonne's projection and the Azimuthal Equal-area projection for a map of the North American continent. Both of these are equal-area projections so that the best way of comparing them is through maximum angular deformation, ω. The origin of both projections is the point with latitude 45° North, 100° West. Isograms for maximum angular deformation are shown for both projections at intervals of ω = 5° and 10°. The shaded patterns refer to the isograms for Bonne's projection. Note that the coastlines are drawn roughly to indicate their approximate location. They do not coincide with their positions on either of these projections accurately and are only an approximate guide to the extent of the area to be mapped. Although this example is for maps and systems of continental dimensions, the same method may be employed for comparison of maps for individual countries. In such cases the isograms would be for values of ω for every degree or even every half degree.
(*Source*: Maling 1973)

TOWARDS AN AUTOMATIC METHOD OF CHOICE OF MAP PROJECTION

Although the method just described was developed using sheets of transparent plastic to represent the overlays, this method of choice is obviously well suited to GIS applications. However, the author has no knowledge whether this particular application has yet been attempted so that there is plenty of scope for further research here.

The only example of the development of an interactive program intended to choose a suitable projection appears to be that by Jankowski and Nyerges (1989) who have tackled the problem in a wholly different fashion. They have proceeded through the medium of existing software packages resulting in the series of programs which they have called the 'Map Projection Knowledge-Based System'. Among the many questions asked in the interactive development of a choice of projection, the user must specify a preference for special property which only distinguishes between conformality, equidistance and equivalence. The default is equivalence and there seems to be, at present, no way of selecting a projection which does not possess one of these special properties. At the stage when Jankowski and Nyerges published their paper, work on the system was still in progress.

REFERENCES

Arthur D W G (1978) Orthogonal transformations. *The American Cartographer* **5**: 72–4

Barton B A (1976) A note on the transformation of spherical coordinates. *The American Cartographer* **3**: 161–8

Briggs D, Mounsey H M (1989) Integrating land resource data into a European geographical information system: practicalities and problems. *Applied Geography* **9**: 5–20

Cocks K D, Walker P A, Parvey C A (1988) Evolution of a continental-scale geographical information system. *International Journal of Geographical Information Systems* **2**: 263–80

Cooper M A R (1974) *Fundamentals of Survey Measurement and Analysis*. Crosby Lockwood Staples, London

Dale P F (1991) Land information systems. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 85–99, Vol 2

Doytsher Y, Shmutter B (1981) Transformation of conformal projections for graphical purposes. *Canadian Surveyor* **35**: 395–404

Fisher P F (1991) Spatial data sources and data problems. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 175–89, Vol 1

ICA (International Cartographic Association) (1973) *Multilingual Dictionary of Technical Terms in Cartography*. ICA, Wiesbaden

Jackson J E (1980) *Sphere, Spheroid and Projections*. Granada, London

Jankowski P, Nyerges T (1989) Design considerations for MaPKBS-map projection knowledge-based system. *The American Cartographer* **16**: 85–95

King C W B (1988) Computational formulae for the Lambert conformal projection. *Survey Review* **29**: 229, 230, 323–37, 387–93

Maling D H (1968) The terminology of map projections. *International Yearbook of Cartography* **8**: 11–65

Maling D H (1973) *Coordinate Systems and Map Projections*. Philip, London

Maling D H (1989) *Measurements from Maps*. Pergamon Press, Oxford

Mather P M (1987) *Computer Processing of Remotely-sensed Images: an introduction*. Wiley, Chichester

Methley B D F (1986) *Computational Models in Surveying and Photogrammetry*. Blackie, Glasgow

Mikhail E M (1976) *Observations and Least Squares*. IEP-Dun-Donnelly Harper & Row, New York

Mounsey H M (1991) Multisource multinational environmental GIS: lessons learnt from CORINE. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 185–200, Vol 2

O'Callaghan J F, Garner B J (1991) Land and geographical information systems in Australia. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 57–70, Vol 2

Richardus P, Adler R K (1972) *Map Projections for Geodesists, Cartographers and Geographers*. North-Holland, Amsterdam

Royal Society (1966) *Glossary of Technical Terms in Cartography*. Royal Society, London

Shmutter B (1981) Transforming conic conformal to TM coordinates. *Survey Review* **26**: 130–6, 201

Sievers J, Bennat H (1989) Reference systems for maps and digital information systems of Antarctica. *Antarctic Science* **1**: 351–62

Snyder J P (1985) Computer-assisted map projection research. *US Geological Survey Bulletin* **1629**. US Government Printing Office, Washington

Snyder J P (1987a) Map projections – a working manual. *US Geological Survey Professional Paper* **1395**. US Government Printing Office, Washington

Snyder J P (1987b) Differences due to projection for the same USGS quadrangle. *Surveying and Mapping* **47**: 199–206

Snyder J P (1987c) Labeling projections on published maps. *The American Cartographer* **14**: 21–7

Sprinsky W H (1987) Transformation of positional geographic data from paper-based map products. *The American Cartographer* **14**: 359–66

Vincenty T (1971) The meridional distance problem for desk computers. *Survey Review* **21**: 136–40, 161

Vincenty T (1989) The flat earth concept in local surveys. *Surveying and Mapping* **49**: 101–2

Williams W B P (1982) The Transverse Mercator Projection – simple but accurate formulae for small computers. *Survey Review* **26**: 205, 307–20

Wray T (1974) The seven aspects of a general map projection. *Cartographica Monograph* **11**, 72 pp.

Wu, Zhong-xing, Yang, Qi-he (1981) A research on the transformation of map projections in computer-aided cartography, *Paper presented at the 10th International Cartographic Conference Tokyo*, 22 pp.