CHAPTER 4

# Coordinate systems, transformations and map projections

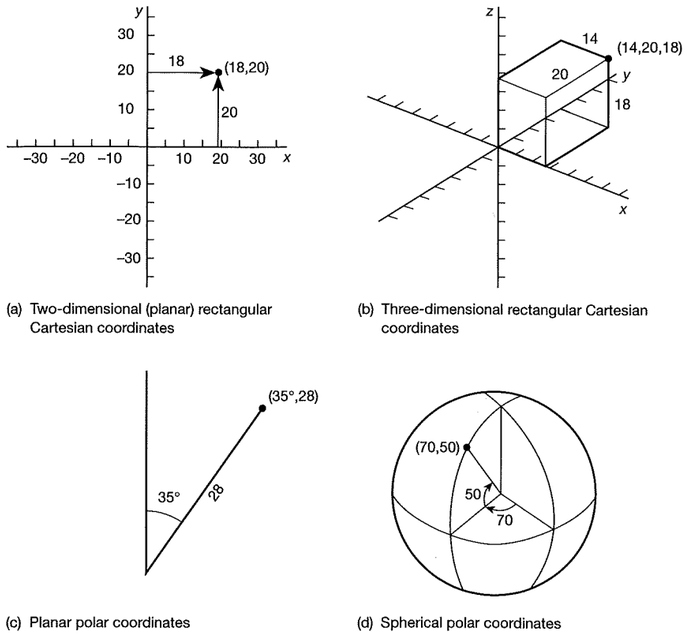
## Introduction

The concept of position as defined by coordinate systems is essential both to the process of map-making and to the performance of spatial search and analysis of geographical information. To plot geographical features on a map, it is necessary to define the position of points on the features with respect to a common frame of reference or coordinate system. Having created such a frame of reference, it also provides a means of partitioning data for purposes of spatial indexing in a database. Thus the coordinate system can be used to guide a search through the database in order to determine which features occur in the vicinity of a point or a region expressed in terms of the coordinates. The coordinate systems that constitute the frames of reference necessary for mapping and searching geographical information allow us to specify position in terms of the distances or directions from fixed points, lines or surfaces (Figure 4.1). In *Cartesian coordinate systems,* positions are defined by their perpendicular distances from a set of fixed axes. The simplest and most familiar example is the case of two straight-line axes intersecting at right angles (Figure 4.1a). In *polar coordinate systems*, positions are defined by their distance from a point of origin and an angle, or angles, which give direction relative to an axis or a plane passing through the origin (Fig 4.1c,d).

Positions on the earth's surface are normally defined by a *geographical coordinate system* consisting of degrees of latitude and longitude. This is a form of spherical polar coordinate system in which two angles are measured with respect to planes passing through the centre of a sphere or approximate sphere (spheroid) representing the shape of the earth (Figure 4.1d). Distance is not specified in the coordinate system but it is implicit, being the radius of the earth at any given location on the surface. Because latitude and longitude refer to positions in 3D space, it is necessary for the purposes of cartography to transform them to a 2D planar coordinate system, or *map grid*. This type of transformation, which is called a *projection*, can be done in many different ways. The principal types of map projection transform from the earth's surface either directly to a plane, or to a cylindrical or a conical surface which, having been conceptually wrapped around the earth, can be unrolled to form a flat surface. When lines of latitude and longitude are plotted on the map they are referred to as a *graticule* (Figure 4.2).

All projections from geographical coordinates on the earth's surface to 2D map-grids involve some sort of distortion. The choice of projection is usually governed by a desire to minimise one or more of the distortions of either angles, linear dimensions or areas. For this reason it is important to appreciate the processes of map projection and the way in which they introduce internal changes in scale which give rise to these distortions.

A consequence of the variety of map projections is that there are numerous map-grid coordinate systems in use, some of which are unique to particular mapping organisations. This means that when compiling databases or maps with data from different sources, it will often be necessary to transform from one coordinate system to another in order to work within a single unifying framework. The use of computers has been most important in facilitating these often




(a) Two-dimensional (planar) rectangular Cartesian coordinates

(b) Three-dimensional rectangular Cartesian coordinates

(c) Planar polar coordinates

(d) Spherical polar coordinates

*Figure 4.1* Cartesian and polar coordinate systems. Cartesian coordinates consist of distances measured relative to fixed axes. Polar coordinates consist of a distance from a fixed origin and an angle or angles representing direction relative to a fixed axis or to a fixed place.

very complex geometric transformations which, when performed manually, may be extremely laborious.

We have already seen in previous chapters that the practice of computer cartography requires working with local coordinate systems which are specific to particular graphics display devices and to particular data acquisition systems. The use of technology for secondary data acquisition may require transformation from digitising table coordinates to geographical or map-grid coordinates, and again from the latter to plotting device coordinates. When, in the course of digitising, it is necessary to compensate for distortion in the source map, (e.g. due to paper stretching), then the transformations become more complicated. Problems also arise when the only coordinate system marked on the map is a non-rectangular graticule of latitude and longitude. In these cases it may be necessary to use interpolation procedures to deduce the map or geographical coordinates of points lying in-between known control points (such as at graticule intersections).

In the following sections of this chapter we start with a discussion of the way in which the shape of the earth is described, before examining the planar and spherical coordinate systems that are used as frames of reference. In the context of this introduction to coordinate systems we review methods for making measurements of length and area, before introducing the basics of simple geometric transformations of translations, scaling and rotation on the plane. The second main part of the chapter is concerned with map projections, including issues of the relationship between the sphere and the surface of projection, the concept of scale and way distortions are introduced in map projections. The following part of the chapter briefly reviews the problems of registering map data by means of rubber sheet transformations when not all projection parameters of existing maps are known. In the final section we introduce a relatively new type of coordinate system based on a tessellation of cells which approximate the shape of the globe.



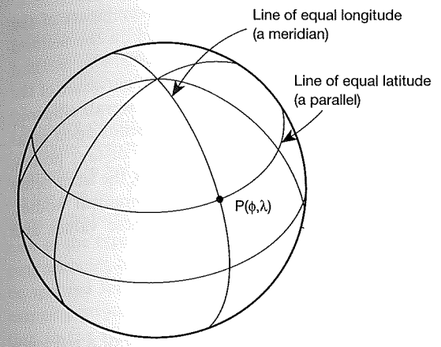



*Figure 4.2* Geographical coordinates are a special case of spherical coordinates in which a point P is defined by an angle of latitude which is measured relative to the plane of the equator (perpendicular to the axis of rotation) and an angle of longitude which is measured relative to a plane of a datum meridian passing through the axis of rotation. The network of lines of equal longitude (meridian) and of equal latitude (parallels) constitute the graticule.

## The shape of the earth

Geographical maps are concerned with representing information that is spatially related to the surface of the earth. To represent the relationships accurately, maps should ideally be scaled down versions of the physical or cultural features of the earth's surface. Since the earth is three-dimensional, this implies creating 3D maps. Apart from the occasional use of globes and physical raised-relief maps, and some specialised 3D viewing systems, this is not generally practicable. It is necessary, therefore, to concern ourselves with the question of how to project the 3D world onto the 2D surfaces which characterise current graphics technology. The process of this projection depends very much upon the shape that we consider the earth to be.

### A flat earth

In small regions, several kilometres in extent, the curvature of the earth's surface departs so little from a plane that it is possible to treat the earth as a flat surface on which local terrain can be measured as perpendicular variations in elevation. The projection of the lateral positions to a planar map is then a trivial matter, requiring only simple scaling of linear dimensions. For more extensive regions, the planar approximation becomes untenable for purposes of accurate locational mapping, since it is impossible to transfer measurements from what, in reality, is a curved surface to a flat one, without introducing distortions such as stretching and tearing.

### Spheres and spheroids

For small-scale maps which cover large areas of the earth, it is appropriate to think of the earth as a sphere of constant radius. From a global point of view this is a very good approximation, as the radius actually only varies by about 10 km to either side of an average value of 6371 km (Maling, 1992). For larger-scale mapping of extensive regions, the spherical approximation is not adequate and account must be taken of the variation from a perfect sphere. Maling (1991) has pointed out that for many GIS applications, using data derived from secondary sources (i.e. digitised maps), the inherent inaccuracies of the data are such that the spherical earth assumption is often quite appropriate.

According to gravitational theory, if the earth were homogeneous in composition, its shape would be an ellipsoid of rotation, generated by rotating an ellipse about its shorter axis. Thus flattening occurs in a north–south direction along the axis of the earth's rotation. In vertical cross-section, the earth's shape would be an ellipse, with the major axis through the equator and the minor axis coincident with the rotational axis. Variations of relief due to mountains and oceans can be regarded as occurring above and below the surface of the ellipsoid, while the direction of gravity would always be normal to the ellipsoid surface, which could be defined as the horizontal at any given location. The shape of the earth represented by this gravitational equipotential surface is called the *geoid*.

The geological composition of the earth is such that there are both major and minor changes in rock density, which give rise to anomalies in the gravity field and hence in the form of the geoid. Satellite observations of the gravity field indicate that it can be represented by a surface that deviates somewhat from an ellipsoid in that it is dented at the south pole and squeezed in northern latitudes to produce something which, when greatly exaggerated, resembles a pear (Maling, 1992). These variations in the form of the geoid are, however, so minor that for the purposes of

large-scale topographic mapping, it is sufficient to treat the earth as an ellipsoid. Because reference ellipsoids used to approximate the earth's shape are so similar to a sphere, they are often described as a *spheroid*, and in this context the terms ellipsoid and spheroid are frequently used interchangeably.

The reference ellipsoid is used to define the *geodetic datum* to which a geographical coordinate system may be linked. The dimensions of this ellipsoid can be defined in terms of the ellipse that generates it (Figure 4.3). The form of an ellipse is defined by the lengths of the semi-major axis ($a$) and the semi-minor axis ($b$). These values can be combined to define the degree of flattening, $f$, also called the ellipticity, oblateness or compression, where

$$f = (a - b)/a$$

The value of $f$ is usually given as a fraction of the form $1/n$, where $n = a/(a - b)$. Several surveys have been made with the intention of estimating the values of $a$ and $b$, and hence $f$, for the earth (see Snyder (1987) and Maling (1992) for lists of examples of officially accepted values). The results have varied somewhat as a function of which part of the world the survey data were obtained from, as is to be expected in view of the slight departures from an exact ellipsoid (as well as error in the survey methods). Several different values of $f$ are in current usage, and most are in the range 1/297 to 1/300, though several larger values of the order of 1/294 are also in use. The measurements for the corresponding values of $a$ and $b$ only vary in general by less than a metre and are of the order of 6378 km and 6356 km respectively (Snyder, 1987).

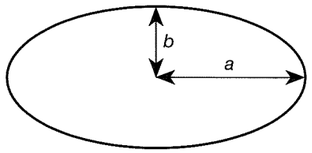



*Figure 4.3* The shape of the earth can be approximated by an ellipsoid, obtained by rotating an ellipse about its minor axis (coinciding with the earth's axis of rotation). In the figure the difference between the semi-major axis *a* and the semi-minor axis *b* is greatly exaggerated compared with the difference between the values used for the earth.

Most of the ground-based measurements of the form of the earth have been based on the method of astro-geodetic arc measurement. The method involves determining various radii of curvature of the earth from which the form of the overall ellipsoid may be deduced. The principle depends upon measuring the distance $d$ between two distant points, along with their angular separation $\theta$, the latter being determined with respect to astronomical bodies. Using the relationship between these values and radius $R$, given by

$$d = R\theta$$

the radius can be found ($R = d/\theta$).

Recent measurements of the shape of a reference ellipsoid have made use of data obtained from satellites. This is the case for reference ellipsoids such as the North American Datum 1983 (NAD 83) and the Geodetic Reference System (GRS) 80 and the World Geodetic System (WGS) 84 (Snyder, 1987).

## Planar rectangular coordinates

When a planar rectangular (Cartesian) coordinate system is adopted for cartographic purposes, the $x$ and $y$, or horizontal and vertical, axes are usually referred to as eastings and northings respectively. The coordinate system itself is called a grid, or map grid, since it is frequently represented by sets of intersecting horizontal and vertical lines, drawn at regular intervals. Examples of such map grids are provided by the Universal Transverse Mercator system and the US State Plane system (both consisting of a set of grids), and the British National Grid (a single grid), all of which calibrate the grids in metres.

Graphical display devices and digitisers employ rectangular coordinate systems, but there is considerable variation in the units of measurement and, to some extent, in the orientation of the axes. At the level of the hardware, graphics display screens consist of 2D matrices of pixels which are individually addressable by means of an integer coordinate system in which unity corresponds to the separation between two adjacent pixels. Hard copy plotters and digitising tables also have a finite resolution, but it is often possible to work directly with units of millimetres or centimetres, and fractions thereof, rather than just pixels.



## Measurements with rectangular coordinate systems

There are obvious advantages with rectangular coordinates for mapping in that there is a direct relationship between coordinate units and distances on the ground. This statement is subject to the limitations introduced by projection distortions, but in the case of the types of map grid used by national mapping agencies, the distortions are sufficiently small to be ignored for many practical purposes. National mapping projections do usually involve some distortion in the representation of area but, taking the British National Grid as an example, the divergence from true areal scales may be no more than 0.1% (Maling, 1991). The width of the British National Grid is, however, only about 9° of longitude. The standard grids of the widely used Universal Transverse Mercator projection (to which the British system is closely related) are 6°. Distortion of area or of distance becomes more significant for more extensive regions.

### Distance measurements

The shortest distance between pairs of points in a rectangular coordinate system is represented by straight lines which can be measured simply using Pythagorus's theorem. For example, given two points with easting and northing coordinates E1 = 302950, N1 = 2550802 and E2 = 315240, N2 = 2561844, the difference in eastings is E2 – E1 = 12290 and the difference in northings is N2 – N1 = 11042. Their distance apart $D$ is therefore

$$D = (12290^2 + 11042^2)^{1/2} = 16521$$

When we wish to measure the length of a digitised line it is then only necessary to perform the same type of calculation for each successive pair of digitised points and to sum the results.

### Area measurements with rectangular coordinates

Measurement of area on a rectangular grid is a relatively straightforward procedure. Assuming that the region whose area is to be calculated is defined by a polygon consisting of digitised points, it is possible to express the area as the sum of the areas of a set of trapezia. A trapezium is a quadrilateral, *abcd* with two parallel sides. Referring to Figure 4.4, in which *ab* and *cd* are the parallel sides and $h$ which is the distance between them, the area $A$ is given by the formula

$$A = h/2\ (ab + dc)$$

Considering the polygon in Figure 4.4, when vertical lines are drawn from each, clockwise ordered, vertex down to a horizontal line beneath the polygon, we can see that the area of the polygon may be expressed in terms of the differences in area between those trapezia in which the uppermost edge is directed rightwards and those in which the uppermost edge is directed leftwards. If we assume that the width of each such vertical trapezium (corresponding to $h$ above) is given by the difference in $x$ coordinates of successive vertices, we will find that some widths are positive and some negative, corresponding to the senses of right and left direction. This leads to the result that some areas calculated by formula $N$ will be positive and some negative. Adding together all trapesium areas then gives the area of the polygon, since the positive areas correspond to the trapezia of the upper edges and the negative ones to those of the lower edges. The polygon area $A_p$ is then given by the formula

$$A_p = \sum_{i=1}^{n-1} (X_{i+1} - X_i)\ ((Y_i - Y_b) + (Y_{i+1} - Y_b))/2$$

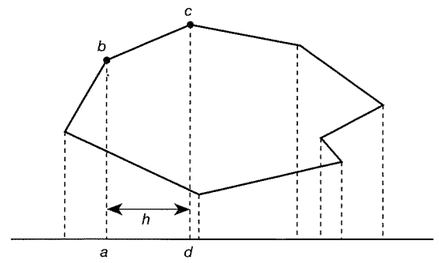



*Figure 4.4* Measurement of the area of a polygon can be performed by summing the positive and negative values found for the areas of the individual trapezia (*abcd*) constructed relative to a horizontal line. The height of each trapezium is found by subtracting successive $x$ coordinates on the boundary, represented by a list of clockwise, or anticlockwise, encoded vertices.

where $X_i$, $Y_i$ are the coordinates of each vertex and $Y_b$ is the $y$ coordinate of a horizontal line. Since $Y_b$ may take any value, including zero, the formula may be simplified to

$$A_p = \sum_{i=1}^{n-1} (X_{i+1} - X_i)(Y_i + Y_{i+1})/2$$

The computation required to find the area can be reduced if we express the formula as follows:

$$A_p = 1/2 \left[ (X_n Y_1 - X_1 Y_n) + \sum_{i=1}^{n-1} (X_i Y_{i+1} - X_{i+1} Y_i) \right]$$

### The centroid of an area

The centroid of an area is a representative point that is usually regarded as being a mean of all locations in the area. A simple approximation to this location could be found by taking the average of the coordinates of all points that defined the perimeter. However, if the density of points along the boundary was variable, this would result in skewing the location of the resulting centroid. A more accurate method is to triangulate the area and find the area-weighted mean of the centroids of the triangles. A triangle centroid is found at the intersection of the lines joining each vertex to the centre of the opposite side. For a concave area, the centroid, found by whichever method, may not be internal. If found to be external (using a point-in-polygon test, as described in Chapter 11), an internal centroid can be found by extending a line horizontally from the initial centroid and finding the centre of an adjacent pair of area boundary intersections.

## Polar coordinates on the plane

An alternative to planar rectangular Cartesian coordinates is the polar coordinate system in which position is defined by the distance $r$ from an origin and the angle $O$ relative to an axis passing through the origin (Figure 4.1c). This type of coordinate system is of use in cartography when plotting certain types of projection in which position can conveniently be defined relative to a single, central point, rather than a pair of axes. It is also relevant in general when it is appropriate to retain a sense of relative direction.

In mathematical usage, polar coordinate angles are conventionally measured anticlockwise from a horizontal axis (Figure 4.1c). In surveying and cartography, the angles are usually measured clockwise from a vertical axis, the angular units being either degrees (0–360) or grads (0–400), where 90° and 100 grads are equivalent to $\pi/2$ radians.

For a given origin, a cartographic polar coordinate of $r,\theta$ can be expressed in rectangular coordinates using the trigonometric formulae

$$x = r\sin\theta, \quad y = r\cos\theta$$

Conversely, the polar coordinates can be expressed in terms of the rectangular coordinates by finding $\theta$ from the relationship

$$\tan\theta = x/y$$

so that $\theta$ can be found using the appropriate inverse tan function on a computer. Knowing $\theta$,

$$r = y/\cos\theta \quad \text{or} \quad r = x/\sin\theta$$

We may also note that

$$r^2 = x^2 + y^2$$

## Spherical coordinates

We have seen that though planar coordinate systems are essential for constructing maps on flat surfaces, they cannot be used for representing extensive regions of the earth without introducing serious distortion in measurements such as distance and area. When high accuracy is not required these problems of distortion can be avoided by the use of a spherical coordinate system. This provides a single, consistent and relatively undistorted reference frame for recording positions and making measurements of the earth's surface. The coordinates can then be projected to a suitable planar coordinate system when a small-scale map of a particular region or aspect of the earth is required.

Any point on the surface of a sphere of given radius can be uniquely defined by the angles which the radius passing through the point makes with two reference planes passing through the centre (Figure 4.5). This is equivalent to a 3D polar coordinate system in



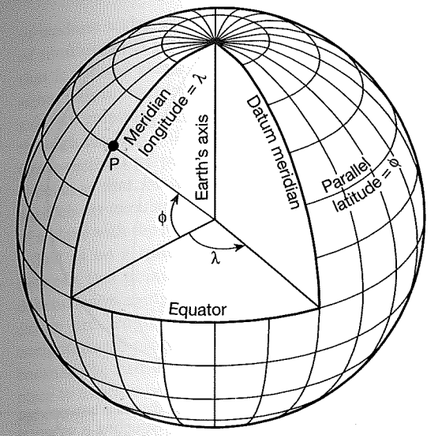



*Figure 4.5* Cutaway view of the earth showing the angular relationship between a point $P$ at latitude $\phi$ and longitude $\lambda$ and the planes of the equator and the datum meridian. Note that $\phi$ is measured in the plane of the meridian which passes through $P$, and $\lambda$ is measured in the plane of the equator.

which the distance from the point to the origin is fixed and hence the locus of all possible positions describes a sphere. On the earth, when it is treated as a sphere, the reference planes of the geographical coordinate system are the horizontal one perpendicular to the axis of rotation, which intersects the surface at the *equator*, and the vertical one which includes the rotation axis and intersects the surface on an arc called a *meridian*, which, for international purposes, passes through Greenwich, London. Angles measured in the vertical, meridianal planes relative to the equatorial plane constitute latitude, while those measured in the horizontal, equatorial plane relative to the plane of the Greenwich meridian constitute longitude. These two angles of latitude and longitude are also described as $\phi$ (phi) and $\lambda$ (lambda) respectively. Although the meridian through Greenwich is the one most commonly adopted as 0° longitude, many national surveys measure longitude relative to a meridian which passes through their capital city.

The word meridian refers specifically to the semi-circular arc formed by the intersection with the earth's surface of any plane which includes the axis of rotation. A single meridian is constituted by an arc that extends from the north pole to the south pole where the word *pole* refers to the intersection of the rotation axis with the earth's surface. For any given meridian there exists an *antimeridian,* which is the arc in the same plane extending around the opposite side of the earth. Angles of longitude are conventionally negative when measured westwards of the zero meridian and positive when measured eastwards. Frequently the sign is omitted, direction being indicated by the specification of east or west.

Angles of latitude are defined either north or south of the equator, where northwards is conventionally treated as positive and south is negative. All points of a particular angle of latitude on the earth describe a circle, the plane of which is parallel to that of the equator. These circles, or lines of latitude, are referred to as *parallels*.

### Great circles and small circles

In a spherical coordinate system, any line between two points must be curved, since it lies on a sphere. An important class of such lines is the circular arcs which result from the intersection of a plane with the sphere. Both meridians and parallels belong to this category. If the plane passes through the centre of the sphere, then the arc is of maximum radius, equal to the radius of the sphere, and it is termed a *great circle*. The shortest distance between any two points on a sphere is given by the length of the great circle arc which extends between them (Figure 4.6). Provided that the two points are not diametrically opposite, only one great circle will fit through them and the shortest distance will be that of the shorter of the two arcs which connect them.

It follows from the description of the meridian given above that all meridians are great circles. Thus all parallels are small circles, with the exception of the equator (which is a great circle). It should be noted that the shortest path between two points of the same latitude will not, therefore, be along the corresponding parallel, unless it is the equator.

### Measurement along any great circle arc

In order to measure distance along any great circle arc, and hence be able to find the shortest distance between any two points on a sphere, it is necessary to make use of spherical trigonometry. Considering

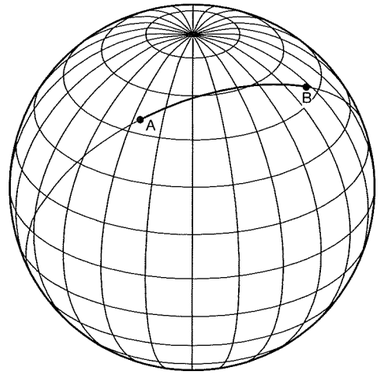



*Figure 4.6* The shortest distance between two points A and B measured on the earth's surface, is given by the length of the great circle arc which extends between them. The great circle arc lies on a plane which passes through the centre of the earth. All meridians are great circles. The equator is the only parallel that is a great circle.

Figure 4.7, it is possible to regard the great circle distance between any two points $A(\phi_a,\lambda_a)$ and $B(\phi_b,\lambda_b)$ as the length of one side of a spherical triangle (i.e. a triangle on the surface of a sphere, as illustrated in Figure 4.8) in which the other two sides are the meridian arcs from the points to the nearest pole. If the latitudes of the two points are $\phi_a$ and $\phi_b$, their angular distance to the pole will be $c1 = 90 - \phi_a$ and $c2 = 90 - \phi_b$, where $c1$ and $c2$ are called the colatitudes. The spherical angle opposite the unknown side is the difference $d\lambda$ in longitudes of the two points where $d\lambda = \lambda_a - \lambda_b$. It is a simple matter, therefore, to apply the cosine formula to derive the unknown angular distance $d$, where

$$\cos d = \cos c1.\cos c2 + \sin c1.\sin c2.\cos d\lambda$$

This formula can also be expressed in terms of the latitudes where

$$\cos d = \sin\phi_a.\sin\phi_b + \cos\phi_a.\cos\phi_b.\cos d\lambda$$

Having found the value of $d$, the angular distance in radians, the arc length $s$ is then given by

$$s = R.d$$

where $R$ is the radius of the earth.

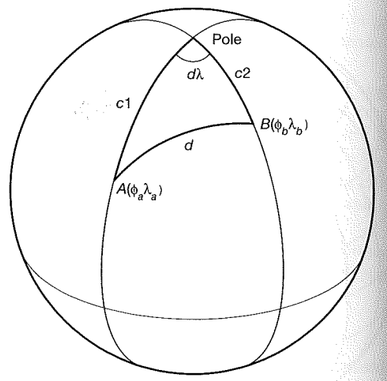



*Figure 4.7* The shortest (great circle) distance between two points A and B can be found by constructing a spherical triangle between the two points and the pole. The angular distance can then be calculated, using the cosine formula, from the colatitudes c1 and c2 of the points, and the spherical angle $d\lambda$, which is the difference between the longitudes of the two points.

## Geometry of the spheroid

It has already been noted that, for the purposes of accurate surveying and mapping of the earth at large

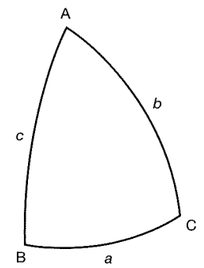



*Figure 4.8* A spherical triangle. Each side is part of a great circle on a sphere. Each angle (A, B and C) is measured between the tangents of the two adjacent sides at the triangle corner. For a given corner, the tangents will lie in the plane which is a tangent to the sphere at that point.



or medium scales, the earth is treated as an ellipsoid of rotation. It is generated by specifying an ellipse that is rotated about its minor (shorter) axis. The minor axis is regarded as coincident with the earth's axis of rotation and thus planar cross-sections through the axis will always be elliptical in shape. Cross-sections perpendicular to the axis are always circular, so that, on the spheroid, the equator and all parallels are circular. Note that no other intersections of a plane with the spheroid will result in circles.

Just as on a sphere, measurements of distances between two points on the surface of a spheroid depend upon using the radius of curvature to calculate arc length. This introduces considerable computational problems for spheroidal calculations because the radius of curvature varies from one place to another on the surface as well as varying at each point, according to which direction it is measured in. Further explanation of the characteristics of the spheroid can be found in Maling (1992) and details of methods for measuring lengths of shortest distances between two points on the surface of a spheroid, i.e. the geodesic, are given in Maling (1989: 548–549).

## Geometric transformations in rectangular coordinate systems

The basic geometric transformations of translation, scaling and rotation are essential requirements for computer visualisation and manipulation of map data. Combinations of these basic transformations are referred to as *affine transformations*. They are needed when changing between coordinate systems and when changing the location, orientation and size of displayed map symbols. The former requirement arises when data represented in 2D or 3D map-grid coordinates must be displayed on a computer device with its own coordinate system and, in the context of data acquisition, when data recorded initially in a local survey coordinate system must be transformed to a standard map grid. If a survey was recorded on a paper map, the latter transformation could include an intermediate representation in digitising table coordinates.

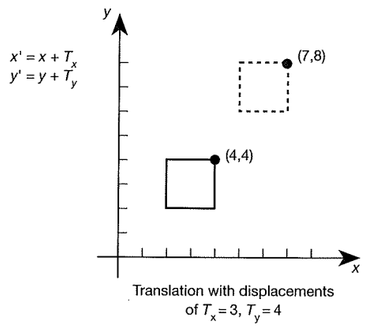



*Figure 4.9* Translation transformation.

### Translation

Translation in 2D (Figure 4.9) moves a point $x,y$ to a new position $x'$, $y'$ by adding components $T_x$, $T_y$, i.e.

$$x' = x + T_x$$
$$y' = y + T_y$$

### Scaling

To scale a point $x,y$ to a new point $x'$, $y'$ we use the two scale factors $S_x$ and $S_y$ relating to scaling in each of the two (or three) dimensions (Figure 4.10). Thus for two dimensions

$$x' = x.S_x$$
$$y' = y.S_y$$

This scaling can be regarded as taking place relative to the origin of the coordinate system. When applied to an object defined by a number of points, the whole object is liable to be displaced, in the sense that the centre of the object will move by an amount determined by the scale factors. If it is required to keep one point of the object fixed, then that fixed point should be moved to the origin before applying the scaling, after which the 'fixed' point should be moved back to its original position. For a fixed point $F_x$,$F_y$ the scaling therefore consists of the composite transformation of $(-F_x, -F_y)$ $(S_x, S_y)$ $(F_x, F_y)$. Hence to

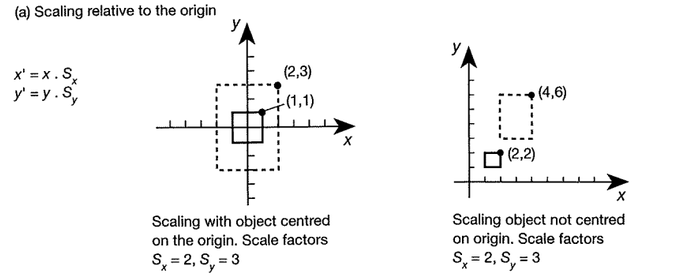



*Figure 4.10* Scaling transformation.

find the transformed point $x'''$, $y'''$ from the initial point $x,y$, we can calculate as follows:

$$x' = x - F_x$$
$$y' = y - F_y$$

$$x'' = x'.\ S_x = (x - F_x)\ S_x$$
$$y'' = y'.\ S_y = (y - F_y)\ S_y$$

$$x''' = x'' + F_x = x.S_x + F_x(1 - S_x)$$
$$y''' = y'' + F_y = y.S_y + F_y(1 - S_y)$$

### Rotation

To rotate a point $x,y$ about the origin by an angle $\theta$ to a new position $x'$, $y'$ the formula is

$$x' = x\cos\theta - y\sin\theta$$
$$y' = x\sin\theta + y\cos\theta$$

The rotation, as defined above, is anticlockwise relative to the origin of the coordinate system (Figure 4.11). In order to rotate an object about an arbitrary origin it is necessary to introduce translations relative to the given point before and after the rotation in a similar manner to that described above for scaling. The latter local rotations are relevant in computer cartography when manipulating individual map symbols such as arrows and items of text.

## Changing between rectangular coordinate systems

To change from one coordinate system to another, one of the coordinate systems must be defined in terms of the other. Assuming, as we are here, that both coordinate systems are rectangular, we need to know a common point that can be defined in both coordinate systems, scale factors in each dimension (i.e. how many units of one coordinate system there are for each unit of the other), and the orientation of



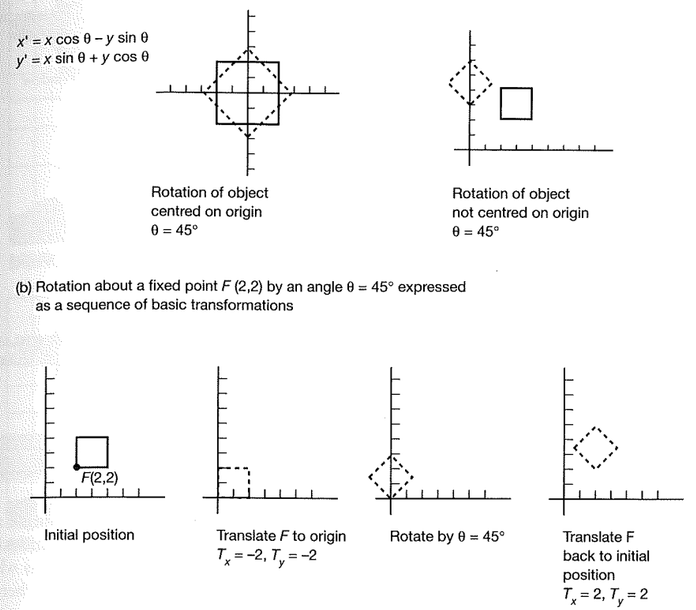



*Figure 4.11* Rotation transformation.

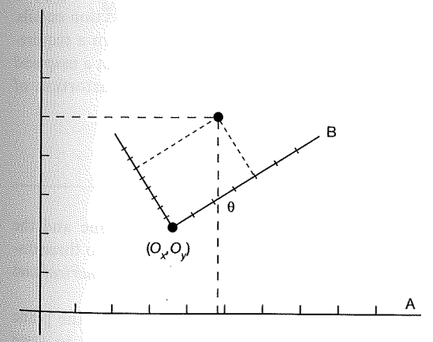



*Figure 4.12* A change in coordinate systems from system A to system B can be achieved by applying to the points in coordinate system A those transformations required to align the coordinate system of B with that of A.

the axes of one system relative to the other. Let us suppose, referring to Figure 4.12, that we wish to transform points in coordinate system A to coordinate system B. The situation is particularly simple if the common point is the origin of B. The transformation consists of applying to the points represented initially in system A those transformations required to align the coordinate system of B with that of A. Referring to Figure 4.12, $O_x$,$O_y$ is the origin of B defined in coordinate system A, $B_{sx}$, $B_{sy}$ are the numbers of units of B per unit of A in the $x$ and $y$ axes, and $\theta$ is the angle between the $x$-axis of B and the $x$-axis of A. The steps are then

1. Translate by $-O_x$, $-O_y$ moving the origin of B to that of A.
2. Rotate by $-\theta$, bringing the axes into alignment.
3. Scale by $B_{sx}$, $B_{sy}$ to make the units of each dimension equivalent.

If the common point used for translation in step 1 is not the actual origin of coordinate system B, but is

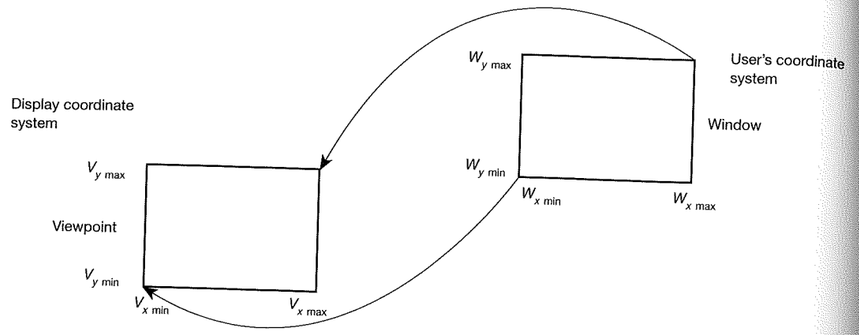



*Figure 4.13* The window to viewport transformation in computer graphics. A change in coordinate systems is required to display a rectangular region of the user's (world) coordinate system, defined by a window, within a rectangular region of a display device, defined by a viewport.

some point $b_x$, $b_y$ in the coordinate system of B, we need to introduce a fourth step to add on the coordinates of this point. The additional step is

4. Translate by $b_x$, $b_y$ to add on the local origin.

A practical example of this change in coordinate systems would be where coordinate system B represents the grid system of a map being digitised on a digitising table represented by coordinate system A. If this was the case then the scale factors and angle might not be specified explicitly, but could be derived from three control points which the operator was required to specify. These points might be the origin of the map sheet and one point on each of the two adjacent corners of the map grid.

Another example of an application of the coordinate system transformation is that of plotting a digital map defined in its own coordinate system on a graphics display device. In computer graphics terminology, the transformation would be defined by a window on the user's coordinate system, i.e a rectangular area defining the region of the map to be displayed, and a viewport defining a corresponding rectangle on the display surface of the device in the graphics display coordinate system. Usually both rectangles are oriented parallel to each other, so that there is no need for the rotation step (Figure 4.13). The initial translation is given by the lower left corner of the viewport. The scalings are derived from the ratios of the sides of the window and viewport and the final translation is given by the coordinates of the lower left of the window.

## Map projections

Map projections involve transforming a representation of the world in three dimensions to a representation in only two dimensions which can be plotted directly on a planar surface. There are many different ways in which the projection can be performed and, in order to understand the variety of map projections that result (Figure 4.14), it is useful to take account of several aspects of the problem. Of particular relevance are (a) the relationship of the planar surface to the global surface and (b) the nature of the distortion which the projection entails. The issue of distortion introduces in turn a requirement to consider variation in scale within a map and its specific association with the distortion of lengths, angles and areas.

### Disposition of the plane of projection

The relationship between the map plane and the spherical or spheroidal surface gives rise to three distinct classes of projection: *azimuthal*, *cylindrical* and *conical* (Figure 4.15).

In the first case, of planar projection, the plane may be regarded as lying flat at a tangent to some point on the globe. This results in an azimuthal projection which takes its name from the particular case in which the plane is a tangent to the north or south



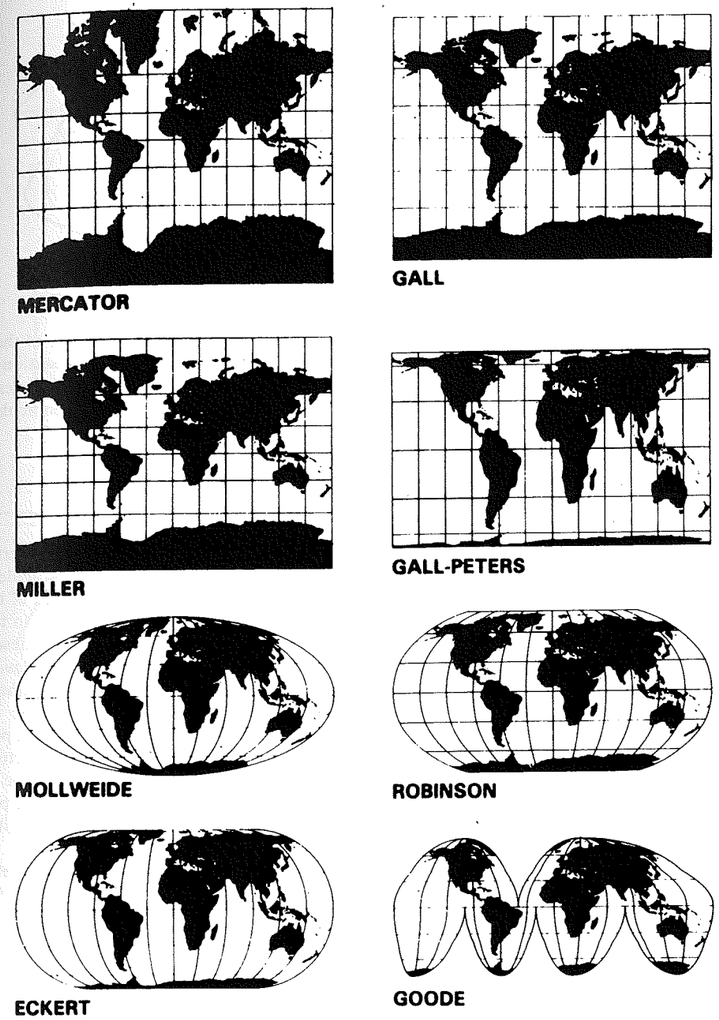



*Figure 4.14* Examples of world map projections. Reproduced by kind permission of the University of Wisconsin Cartography Lab.

pole from which meridians will radiate according to their respective azimuths. In the second, cylindrical, class of projection the plane is derived from a surface wrapped cylindrically around the globe, touching it along a single great circle. Note that unwrapping the cylinder into a genuine planar surface can be done

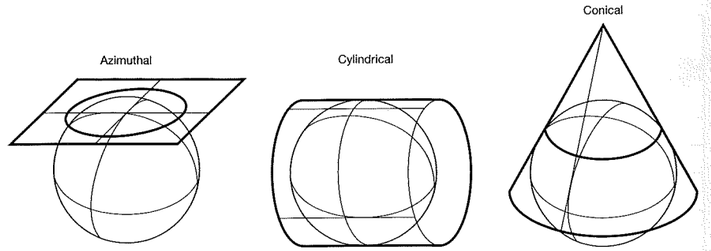



*Figure 4.15* The three main classes of map projection.

without introducing further distortion. In the third class of projection, the conic, the map plane is derived from a surface that is wrapped around the earth in the shape of a cone touching the globe along a small circle or an ellipse. Projection surfaces such as the cone and cylinder, which can be unwrapped without distortion to form a plane, are called *developable surfaces*. It should be borne in mind when referring to planar, conical and cylindrical projections that ideas of the tangential plane, cylinder and cone are geometric concepts which are intended to help envisage the projection transformation. The projections are defined by mathematical functions which transform points on the globe onto the map surface along projection rays emanating in various ways from the globe to the map.

Each of the three forms of projection has modifications in which the projection surface intersects the earth rather than touching it tangentially at a point or along an arc (Figure 4.16). When the plane, cylinder or cone intersects the earth, it is described as *secant*. The secant plane intersects a spherical globe along a single circular arc. Both the secant cylinder and the secant cone intersect along two parallel small circles. If the ellipsoidal shape of the earth is taken into account, the arcs of intersection will only be exact small circles if they lie in planes parallel with the Equator and hence perpendicular to the rotational axis.

## Aspect

The appearance of a map projection, in terms of the form of the latitude–longitude graticules, varies according to the orientation, or aspect, of the plane, cylinder or cone of projection relative to the globe. The aspect of a projection is usually described as being either *normal*, *transverse* or *oblique* (Figure 4.17). The normal aspect of the azimuthal projection refers to the situation when the plane of projection is located at one or other pole, perpendicular to the earth's axis. It is characterised by radiating lines of longitude and concentric circles of latitude. The normal aspect of a conical projection has a similar pattern of latitude and longitude, and arises when the cone touches or intercepts the globe on a line or lines of latitude, corresponding to the

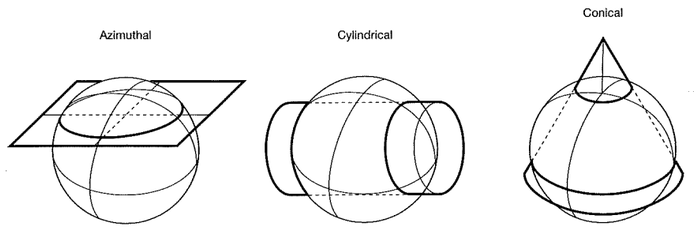



*Figure 4.16* Secant projections are ones in which the plane, cylinder or cone of projection intersects the globe, as opposed to touching it tangentially.



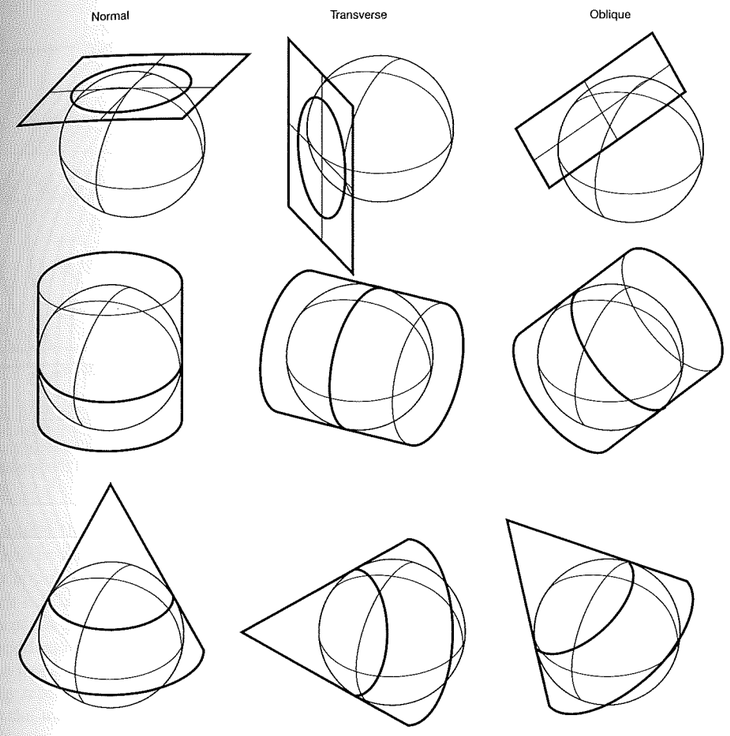



*Figure 4.17* Aspects of azimuthal, cylindrical and conical projections. In the normal aspects of the azimuthal and conical projection, the plane and cone respectively may be positioned at the north or south poles. In the transverse aspects, the axes of the cone and cylinder, and the perpendicular to the plane may take on any horizontal (equatorial) orientation.

axis of the cone being parallel to the earth's axis. The transverse aspects of both azimuthal and conical projections occur when the orientation of the plane and cone is at 90° to that of the normal aspect. The appearance of the resulting maps are distinguished by the equator being horizontal while one, central meridian is vertical. Oblique aspects refer to all possible intermediate orientations of the plane and cone.

Considering the case of cylindrical projections, the normal aspect refers to the situation in which lines of latitude are horizontally orientated and lines of longitude are vertically orientated, though these lines may be curved, depending upon other properties of the projection. The normal aspect of cylindrical projection corresponds to the cylinder being orientated north–south parallel to the earth's axis. In the transverse aspect, the cylinder is orientated east–west and, when tangential, touches the globe along a great circle. The graticule includes a vertical central meridian and a horizontal equator. Oblique aspects of cylindrical projections refer to all other orientations of the cylinder relative to the globe.

Figure 4.18 illustrates the way in which the widely used Universal Transverse Mercator (UTM) projection system specifies 30 orientations for the horizontal cylinder, providing 60 UTM zones, each of 6° width (one on either side of the cylinder). The relationship between an individual zone and the corresponding grid system is illustrated in Figure 4.19.

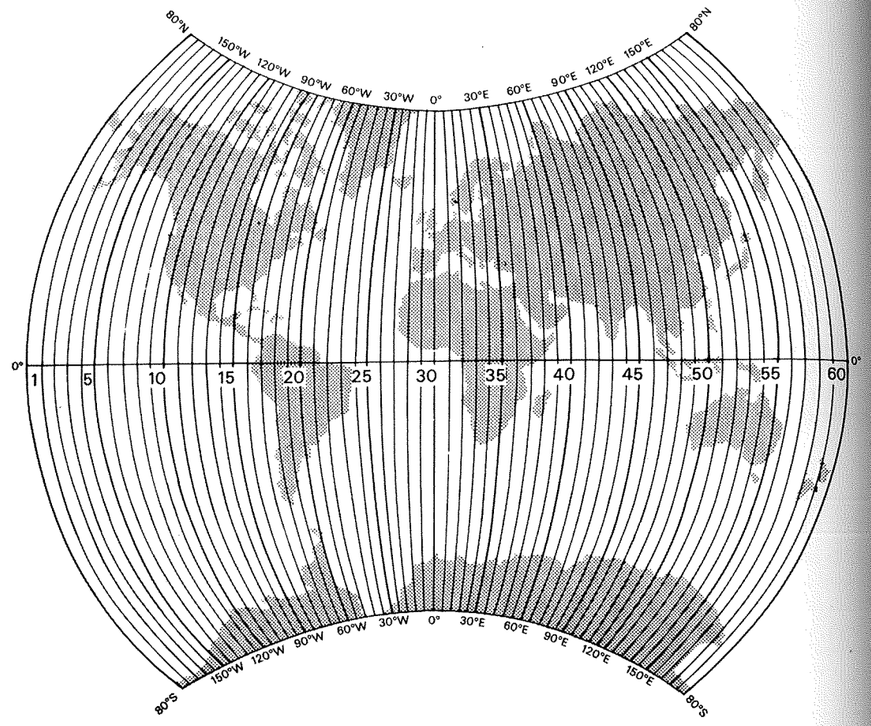



*Figure 4.18* The UTM projection system employs 30 orientations of a cylinder, giving 60 zones each 6° wide. After Maling (1992).

## Concepts of scale

Geographical maps can be thought of as scaled graphic representations of physical or abstract features of the earth. The map reader usually assumes that dimensions on the map can be related to their true dimension in terms of a scale value which may be expressed by the ratio between map dimension and actual dimension. When the scale value is given as a fraction in which the numerator is 1, it is called the *representative fraction*. If the representative fraction is relatively large, such as 1/1250 for a detailed urban survey, the map is referred to as *large scale*, whereas maps covering much larger areas in which the representative fraction is correspondingly smaller, such as 1/500 000, are referred to as *small scale*. In general, the meaning of large, small and medium scale will depend upon the conventions of particular applications.

The direct relationship between map dimension and actual dimension, implied by the representative fraction, is in practice somewhat misleading. This is because the distortion involved in transferring dimensions from the curved surface of the earth to a planar map prevent the possibility of a constant scale throughout the map.



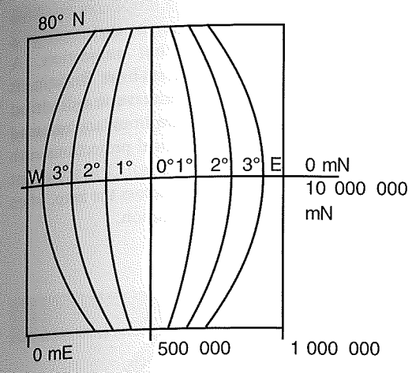



*Figure 4.19* Each 6° wide UTM zone corresponds to a pair of map-grid systems for the north and south hemisphere respectively. Each grid extends from 0 to 10 000 000 in the north direction and 0 to 1 000 000 in the east direction.

Conceptually the production of a map requires that the earth be scaled down to a *generating globe* of a size that is compatible with the size of map plane onto which it is to be projected. The ratio between the radius of the generating globe and the radius of the real world equivalent is called the *principal scale*, and it is to this scale that the representative fraction refers. The representative fraction can only apply exactly to the scale at one or two specific points or lines on the map, which are positions of zero distortion. These locations are those of contact or intersection between the projection plane, cone or cylinder and the globe. Away from these positions of zero distortion, the scale varies in a manner that depends upon the type of projection. The variation of scale leads to the notion of particular scale, which refers to the scale in a specified direction at a specified point on the map.

## Distortion

Understanding the way in which scale varies on map projections is equivalent to understanding the nature of the distortion that all map projections contain. One means of analysing distortion, devised by Tissot, is to examine, for corresponding points on the map and the globe, the way in which an infinitely small circle on the generating globe changes shape when plotted on the map. For positions on the map where there is zero distortion, the circle will remain a circle of the same radius. In all other cases, the circle becomes transformed in terms of either its size, its shape, or both. In the general case, the circle becomes an ellipse, the semi-major and semi-minor axes of which we may refer to as $a$ and $b$ respectively. The orientation of the two axes bears a special relationship to the globe, in that there is normally one pair of perpendicular directions at any given position on the globe which remain perpendicular after projection and assume the orientation of the axes of the ellipse. Tissot called the ellipse an *indicatrix* and its axes are called the *principal directions* at the given point. If the original circle is defined as being of unit radius, then the two semi axis dimensions, $a$ and $b$, are equal to the maximum and minimum particular scales for that point on the map.

The relationships between the values $a$ and $b$ can be used to characterise the distortion properties of a projection. The product of $a$ and $b$ is related to the area of the ellipse and gives rise to a quantity known as the area scale $s$ (also known as $p$) where

$$s = a.b$$

If the area of the ellipse remains the same as the original circle then $s = 1$. Maps in which this property holds are called *equal-area projections*, since they involve no distortion of area between the globe and the map. Maintenance of equal area is accompanied by considerable variation in the individual values and orientations of $a$ and $b$.

The consequence of any departure from circularity of the original circle is the distortion of angles. Angular distortion can only be avoided by keeping $a = b$, i.e. making the ellipse circular. Maps in which $a = b$ at all points are known as *conformal projections*. It is important to realise that the two properties of equal area and conformality are always exclusive on a projection from a globe.

On maps which include angular deformation, it is of interest to be able to monitor the degree of deformation at different points on the map. For every point, there is a maximum value of angular deformation $\omega$, given by the change in angle between two directions on the globe, each of which has undergone

the maximum individual deflections for that point. The value of ω is given by the formula

$$\sin\frac{\omega}{2} = \frac{a-b}{a+b}$$

## Choosing a map projection

When deciding on the most appropriate map projection for a particular purpose, one important choice is between equal area and equal angle (conformal) projections. When representing a very large area on a small-scale map, it will usually be the case that an equal area map is preferable, as it will result in a realistic representation of the relative size of different regions. Thus projections commonly used for world maps, such as the Mollweide and Goode projections, are equal area (Figure 4.14). Note that the Mercator projection, which is now rarely used for world mapping, results in gross distortions of the relative size of continents (Figure 4.14). For high- accuracy maps on which measurements of angle may be made, an equal angle projection is appropriate.

When choosing between planar, conic and cylindrical projections, and their respective aspect, the issue of distortion should be considered. It was pointed out earlier that the point or line of true scale on a map projection coincides with the point (or lines) of contact of the projection surface and the globe. Distortion increases with increasing distance from these locations. To reduce the maximum degree of distortion and to provide an even spread of distortion, the point or lines of true scale should be located as centrally as possible.

Thus for mapping a circular region, such as the whole of the Antarctic, a normal planar (aximuthal) projection is appropriate, since distortion will increase equally in all directions from the centre outwards. For an elongated region either a conic or a cylindrical projection may be appropriate, the objective being to locate the true scale line or lines axially relative to the elongation. For an east–west extended map, a normal conic projection is appropriate since distortion will be at a minimum along lines of latitude.

The use of a secant projection introduces two lines of true scale and hence reduces the distortion in a north–south direction for that projection. A north–south oriented region, such as Britiain, is appropriately mapped with a transverse cylindrical projection, since the true scale lines are oriented north–south. Transverse cylindrical projections such as the UTM and the British National Grid are secant and hence reduce the distortion across the map, when compared with a tangential projection.

## Analytical transformations

Transformations from geographical coordinates to the grid system of a particular projection can be specified precisely using mathematical formulae. In a few cases the formulae are relatively simple, particularly when the earth is assumed spherical. Thus for the Mercator projection

$$x = R.\lambda$$

$$y = R \ln(\tan(\pi/4 + \phi/2))$$

where $x$ and $y$ are grid coordinates, $\lambda$ and $\phi$ are longitude and latitude, $R$ is the radius of the scaled sphere, and ln is the natural logarithm (i.e. to the base e). The equations are called the *forward solution*. The reverse transformation from grid coordinates to geographical coordinates is called the *inverse solution* and, in our simple example of the Mercator projection, is given by

$$\phi = \pi/2 - 2\tan^{-1}(e^{-y/R})$$

$$\lambda = x/R + \lambda_o$$

where e is the base of natural logarithms and $\lambda_o$ is the meridian passing through the origin of the geographical coordinate system.

For other map projections, forward solutions are often more complicated than the above example and become especially so when the transformation takes account of the non-spherical form of the earth. The inverse solutions often require an iterative form of solution. Details of the solutions for many important transformations are found in Snyder (1987).



## Rubber sheet transformations

The use of the various well-defined formulae and procedures for converting back and forth between geographical and grid coordinates depends upon a clear definition of the projection in use. It is quite common, however, to encounter maps in which inadequate details of the projection are provided. If this is so then it may sometimes be necessary to guess appropriate projections and hope for the best. An alternative approach is to make use of so-called

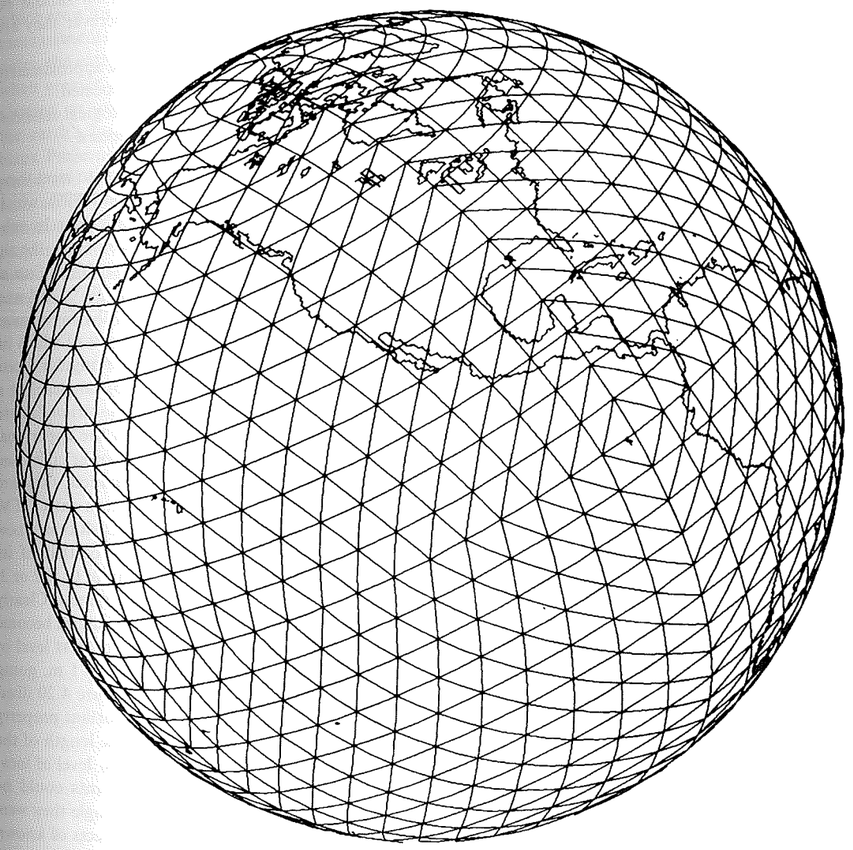


*Figure 4.20* An example of a global tessellation, based on triangular faces. From Goodchild and Shiren (1992). Reproduced with permission from National Center for Geographic Information and Analysis.

rubber sheet transformations, which provide a means of converting a poorly defined coordinate system to a well-defined system, on the assumption that it is possible to identify some control points, the location of which is known in both coordinate systems. Such control points are likely to correspond to the location of features which can be clearly recognised in maps represented in both coordinate systems.

Rubber sheet transformations are also of use in situations other than when a conventional map projection is poorly defined. One important application is for registering satellite imagery with geographical or grid coordinate systems, the problem being that the geometry of the projection from the surface of the earth to the satellite's imaging system may not be precisely defined. Another application arises in the context of map digitising when adjacent or overlapping map sheets are found not to match up precisely, due to errors in digitising and due to distortion in the paper on which the map is drawn.

One technique for rubber sheet transformations makes use of polynomial equations to relate the coordinates in one map to those in the other. Rubber sheet transformations generally need to account for differences which are non-linear and thus require higher-order polynomials. Another approach, which is designed to ensure that the control points transform precisely from one coordinate system to the other, makes use of a triangulation scheme in which the control points are triangulated in the two maps and each equivalent triangle is transformed individually, with points internal to the triangles being interpolated in a linear manner (White and Griffin, 1985).

## Global hierarchical tessellations

Interest in the development of global databases has led to efforts to design locational coding schemes (or geocodings) which allow spatial phenomena to be studied at different levels of detail in a consistent fashion across extensive regions of the earth. We have already seen that conventional projections can only retain simultaneously approximations to properties of equal area and a lack of shape distortion across limited areas. Following Goodchild and Shire (1992), we can identify four desirable properties of a global coding scheme for recording properties of the earth's surface in terms of finite spatial elements (or areal units):

1. The scheme should be hierarchical with elements at each level being subdivisions of elements at the next higher level.
2. Elements at any level of resolution should be approximately the same size, wherever they are located on the globe.
3. Elements at any level should all be approximately the same shape, wherever they are located.
4. The scheme should preserve topological relationships correctly, particularly adjacencies.

One of the simpler schemes which aims to meet these objectives is that described by Dutton (1989), which represents the globe in a hierarchical fashion by subdividing the triangular facets of an octagon, the six vertices of which are located on the earth's surface at the north and south poles and at 0°, 90°, –90° and 180° longitude. An octagon is one of the five regular polyhedra or Platonic solids, the vertices of which lie on a sphere. In Dutton's scheme, the triangular facets are divided recursively into four subtriangles. For a triangle with a horizontal base, the subtriangles are numbered 0 for the central one, 1 for the upper one, and 2 and 3 for the lower left and lower right one respectively. As triangles are subdivided, they are allocated numeric codes which Dutton calls QTM (quaternary triangular mesh) codes. Each time a new triangle is created its code consists of that of its parent with the addition, on the right, of 0, 1, 2 or 3, depending on its location within the parent. Clearly with each level of subdivision the triangles become smaller, and Dutton shows that at the 21st level of subdivision their size is approximately 1 m, going down to 17 cm at the 24th level (Figure 4.20 illustrates a level 4 subdivision). An important property of the QTM codes is the fact that the length of the code can be used to imply a particular level of locational accuracy. Thus individual points could be allocated the code of the smallest triangle they were known reliably to occupy. Similar objects of known areal extent could be allocated to the triangle that completely enclosed them.



## Summary

Much of the data used in GIS are two dimensional, having been derived from maps which are based on projections from the 3D world to a planar map sheet. Locations on the earth's surface can be defined independently of a map by means of the geographical coordinates of latitude and longitude, which are a form of polar coordinates. When making measurements in terms of the geographical coordinates, the earth is assumed to be either a sphere or, when greater accuracy is required, an ellipsoid of rotation, or spheroid. These mathematical models of the earth's form are used when projecting to 2D maps. Projections are classed broadly as cylindrical, planar and conical, according to the type of surface onto which the projection is made, and as normal, transverse and oblique according to the aspect of the projection surface. All map projections introduce some form of distortion, which leads to a further categorisation into equal-area maps, which preserve measurement of area, and conformal maps which preserve angles in the projection. Computing technology has resulted in a recent interest in the possible use of global tessellations of the earth's surface, which provide a discrete, and in some cases variable scale, of cell-based locational referencing in three dimensions.

## Further reading

A more detailed account of much of the material in this chapter concerning map coordinates, the form of the earth and map projections can be found in Maling (1989, 1992) and Snyder (1987), which have been referred to in course of the chapter. For a more extensive coverage of the transformations used in computer graphics, see Rogers and Adams (1990), in particular Chapters 2 and 3. Tobler has published several papers on the transformations used in cartograms, whereby the scale of a map is modified locally in proportion to a mapped variable such as population (see for example Tobler, 1979, 1986). Nyerges and Jankowski (1989) have developed an expert system for helping in selecting map projections.