Field Techniques Manual:
GIS, GPS and Remote Sensing

- <u>Section A: Introduction</u>

Chapter 2: The Geographical Framework

2 The Geographical Framework

One key feature of a GIS allows different types of data – whether from paper maps, GPS, satellite images or aerial photographs – to be integrated and overlaid: and that is the common geographical reference system, or the ‘geographical framework’. Once a piece of information (say the species of a tree or the height of a mountain) has been given a co-ordinate within a commonly recognised reference system (say from a GPS or from a map), it can be added to a GIS. This ability to integrate geographical data underlies all GIS functions, whether for mapping, navigation or more complex analysis. Therefore GIS users need an understanding of how geographical reference systems work. This perhaps applies particularly to expedition GISers, as they deal with data from a variety of sources. From being a traditionally specialist subject, geographical reference systems are becoming common currency for anyone using a GPS or GIS. Expeditioners often come face to face with co-ordinate issues only when things go wrong – layers don’t overlay, or GPS readings don’t match the map – so it is best to be prepared with an understanding of the underlying principles set out here.

In practice, as this chapter explains, there is no one single reference system always used world-wide, but a multitude, according to variations in the shape of the earth and the purposes of the GIS. However, provided that co-ordinates can be transformed between different systems, data can for most purposes be regarded as having a common referencing system. Most current GIS software allows near-instantaneous transformation between referencing systems, but it remains essential to know which system each of your data layers uses. Also note that a co-ordinate reference system is entirely independent of the nature of the data using it; raster, vector and any another spatial data types all ‘inhabit’ the same framework. (Raster is a grid-based data structure, while vector data use co-ordinate geometry; these are explored fully in Chapter 3.)

As examples, here are some situations where you will need to know which system to use:

- Telling a GPS how to display its co-ordinates;
- Saving data downloaded from a GPS;
- Digitizing features from paper maps;
- Fitting an aerial photograph or satellite image to a co-ordinate system;
- Importing data provided by others;
- Deciding on a suitable projection for cartographic output;
- Documenting data you share with others.

This chapter is in three sections: (i) the principles of how geographical reference systems are constructed, the longest section; (ii) details of one particular co-ordinate system, UTM, both as an example and because it is commonly used worldwide; and (iii) how to choose a suitable system for your GIS.

2.1 Geographical reference systems

This section covers the following four stages in turn:

- Find a regular mathematical shape that approximates to the shape of the earth, in the form of an *ellipsoid*. Once its position is defined in relation to the 'actual' earth, it provides a reference surface against on which measure positions – this is a *geodetic datum*.
- Measure positions of our points of interest – whatever they might be – using two angles, *latitude* and *longitude*, in relation to the datum.
- Transform our positions from the curved surface of the ellipsoid onto the flat surface of a map by means of a *map projection*.
- Finally, create a grid on the projected map to provide a convenient rectangular *co-ordinate system*.

Some aspects are considerably simplified here, but the aim is to provide enough detail to start working with GIS and GPS co-ordinate systems. Iliffe (2000) and Jones (1999) are recommended as being more comprehensive yet easy to read, while Maling (1992) is a definitive and very detailed text.

2.1.1 A reference model of the earth (the geodetic datum)

A ruler measures one-dimensional lengths from its zero mark: in other words, it provides a known reference point, also called a *datum*. In just the same way, when we want to establish positions of features on the earth's surface, we need a three-dimensional reference system against which to make measurements. In this case, the model is referred to as a *geodetic datum*.

It is desirable to have a model that fits well to the shape and size of the earth's surface, as that is where positions are measured. The science of determining the shape and size of the earth is known as *geodesy*, and the whole topic of geodesy would be a great deal simpler if the earth was exactly spherical; we would then be able to use a sphere of a certain radius as the geodetic datum for accurate positioning anywhere in the world. (Indeed, a spherical datum is used for some mapping at a global scale, but for expedition fieldwork and most other GIS applications we need greater accuracy.) Two factors mean that we need to refine the simple spherical model: the flattening of the earth at the poles, and irregularities in the shape of the earth.

2.1.1.1 Ellipsoidal datums for a flattened earth

The spin of the earth gives our planet a slight flattening at the poles and a bulge around the equator. Its radius at the equator is approximately 6,378 km, and at the poles is about 6,357 km, a difference of 21 km, or 0.33%. This shape is still relatively simple to model mathematically, using an ellipsoid. This is the figure formed by rotating an ellipse around its short axis (Figure 2-1).

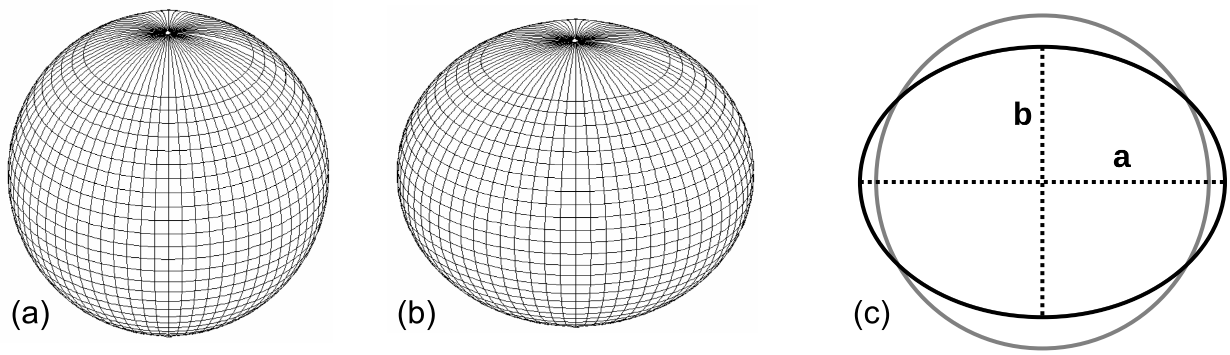


Figure 2-1 (a) A sphere, a relatively inaccurate approximation to the shape of the earth; (b) an ellipsoid, a better approximation; and (c) a cross-section of an ellipsoid, with a circle for comparison, showing the dimensions commonly used to describe an ellipsoid: a = longer radius, or semi-major axis, b = shorter radius, or semi-minor axis.

An optional note on terminology:
'Ellipsoid' and 'spheroid' are usually used interchangeably in geodesy. We have used 'ellipsoid' here as it more readily suggests the shape. However, in the field of geometry, an ellipsoid is a sphere that has been squashed both downwards and sideways – each of its three axes can be of different lengths. In a spheroid, by contrast, only two axes are different (Figure 2-1c), while in a sphere, all axes are of course the same. So while it is not wrong to use the term ellipsoid in this context, it would arguably be more accurate to use spheroid.

2.1.1.2 The 'lumpy' earth, or geoid

As well as being slightly flattened, the earth's 'average' surface level is slightly irregular: even if the planet was entirely sea-covered, its shape would be a slightly lumpy ellipsoid, rather than a smooth mathematical one. Again, the variations are slight, but in terms of positioning, they are significant. This is mainly due to variations in the density of different parts of the earth's mantle and crust. Instead of acting uniformly towards the earth's centre of mass, there are local variations in the direction of the earth's gravitational force, and the result is a slight deformation of surface features.

This irregular surface is called the *geoid*. More specifically, the geoid is a surface of equal gravitational potential. There are of course an infinite number of such surfaces closer or further to the earth, with higher or lower gravitational potential, but the earth's geoid is taken as the surface that coincides on average with the surface of the oceans (Figure 2-2).

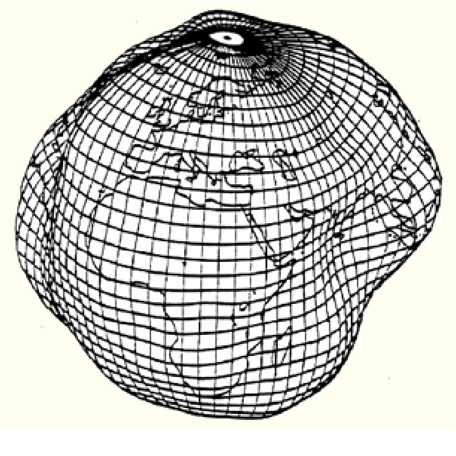


Figure 2-2 An exaggerated view of the geoid, a hypothetical sea-level surface covering the globe. In this figure, the separation between an ellipsoid surface and the geoid surface has been exaggerated by a factor of 15,000. The actual separation ranges from c. -100 m just below India to c. +70 m in the North Atlantic and near New Guinea.

2.1.1.3 Geodetic datums

Because of the irregularities in the shape of the earth, surveyors since the 18th Century have found that while one ellipsoid can provide a good fit to the geoid in one region of the world, a different ellipsoid will be needed in another region. As a result, a multitude of different geodetic datums has been developed. Each datum comprises two elements: (1) an ellipsoid of given shape and size, and (2) a definition of how the ellipsoid is positioned and oriented in relation to the geoid.

Historically, a local datum would be defined first by intensive trigonometric survey work on the ground then by complex calculations to find an optimal solution to fitting an ellipsoid – an astonishing achievement in pre-computer days. For local datums, the positioning of the ellipsoid is typically specified in relation to a particular marked survey point on the earth's surface, known as the initial point, whose position was accurately established by astronomical observations (Figure 2-3). Each of these local datums gives a good result locally, but is unlikely to produce a close fit in other parts of the world. Because of this, a plethora of datums has emerged, each one suited to a certain region, and often revised as surveying of the area improved.

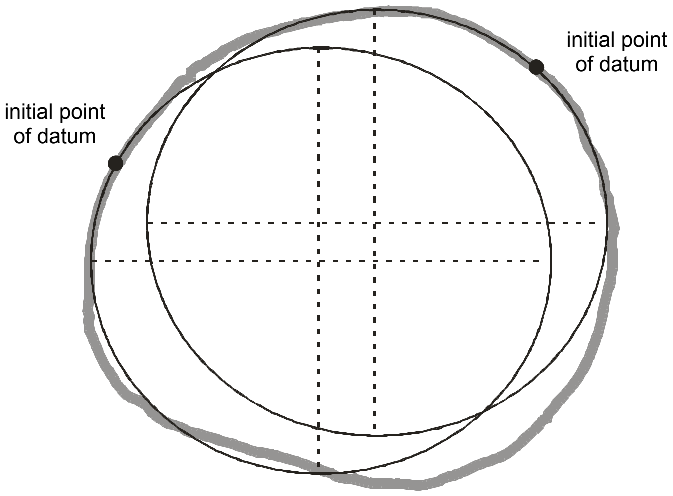


Figure 2-3 A cross section through the earth along the 0°/180° meridian. The grey line shows the mean sea-level surface, or geoid, overlain with two ellipsoids used by two different datums. Each datum is designed to provide a good local fit to the shape of the geoid, with its position defined by its initial point.

In many parts of the world frequented by expeditioners, the best maps are those produced from the late 19th Century up until the 1960s – before the advent of satellite surveying – and being of this era, are based on local datums. In much of eastern and southern Africa, for example, the 'Arc 1960' datum is used (itself a revision of Arc 1950); this is based on the 'Clarke 1880' ellipsoid and is fixed to an initial point in South Africa. The early dates that occur in many ellipsoid names are evidence of the success of the survey and calculation techniques developed in the 19th Century.

More recently, satellite-based measurements of the earth have made it possible to create *geocentric* ('earth-centred') datums. Instead of relating the ellipsoid to a point on the surface, a geocentric datum uses an ellipsoid whose centre coincides exactly with the centre of mass of the earth. The result is a reference system that provides a relatively good fit for all parts of the earth's geoidal surface (Figure 2-4).

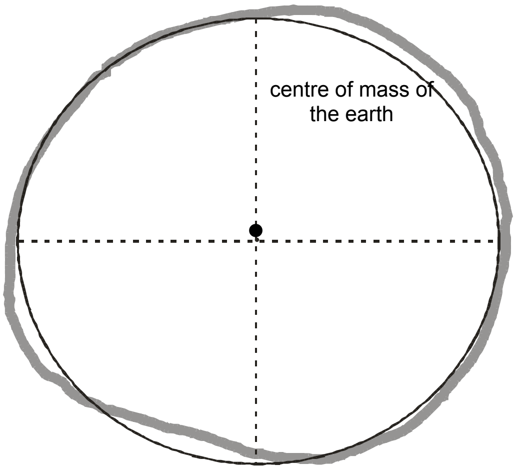


Figure 2-4 The geoid with the WGS84 ellipsoid. Undulations in the geoid have been exaggerated by a factor of 10,000 to make them conspicuous, and the flattening of the earth has also been exaggerated.

Such a datum comprises details of the shape and size of the ellipsoid and its orientation in relation to the earth's centre. This is of particular relevance to fieldworkers, as the GPS system uses a geocentric datum, the World Geodetic System 1984, or WGS84. Indeed, the separations between the ellipsoid and the geoid surfaces given in Figure 2-2 are derived from WGS84: the fit varies from -100 m to +70 m, a relatively tiny fraction of the size of the earth. Chapter 11 gives further details about how this difference is accounted for when measuring elevations with a GPS, as well as the difference between the geoid and the actual earth's surface (Figure 2-5).

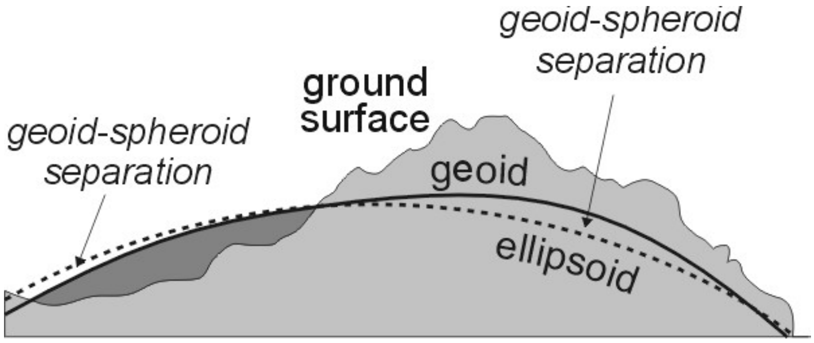


Figure 2-5 How a mathematically-defined ellipsoid surface might fit to the geoid; and the relationship between the geoid (defined as mean sea level around the world) and actual ground topography.

Because of its accuracy, global applicability and use by GPS, WGS84 has become widely used as the 'standard' datum for much geographical information. If all mapping efforts were to re-start from scratch now, almost all GIS and spatially-referenced data would use a system such as WGS84. However, the huge number of older maps and datasets means that many different datums are still in use, and you may decide to use one of the older ones for your GIS – for example, to match local maps. More advice is given later in this chapter. However, it is still vital to be able to transform co-ordinates between datums – for example, from GPS readings that use WGS84 to a GIS of a national park in Tanzania that uses Arc 1960. To do this, conversion software uses a table of known mathematical relationships between WGS84 and other datums. These tables specify the size and shape of the ellipsoid being used, and its position and angle – in three dimensions – in relation to the WGS84 datum. Knowing these parameters, datum conversions can be performed.

One confusion sometimes arises in relation to datums. If you stand at one point, without moving, you have a different co-ordinate depending on which datum you use. On reflection, this is hardly surprising: your measured position depends on where you measure from. By switching to a different datum, your frame of reference is different. The likely errors are significant in relation to the accuracy often needed for fieldwork, ranging from 10s up to 100s of metres. So, a rule of GIS is to make sure that you always (1) know which datum you are using for measuring positions, and (2) state which datum you are using when using, publishing or distributing positional data.

2.1.2 Positions on the ellipsoid: latitude and longitude

Positions on an ellipsoid can be specified with two angles: how far round (longitude) and how far up or down (latitude). Both angles are defined with reference to ‘zero’ positions. Conventionally, a circle passing through Greenwich, London, defines zero longitude, with positive values starting at 0° and increasing eastwards to 180°, and negative values ranging from 0° westwards to 180°. The circle on the plane of the earth’s rotation – the equator, also the widest section of the ellipsoid – defines zero latitude, with values increasing northwards to 90°N and decreasing southwards to 90°S.

For example, the RGS-IBG in London is approximately at 0°10’W, 51°31’N, using WGS84 datum. Simply to say that “the RGS-IBG is at 0°10’W, 51°31’N” is incomplete; for accuracy, we need to know the datum too.

The way in which latitudes and longitudes are written on paper maps and documents differs in several ways from their digital representations. These can easily create confusion, so are explained here.

2.1.2.1 Units: degrees, minutes and seconds

Latitude and longitude, being angular measures, are normally expressed in degrees. To provide greater precision, several different sub-divisions of a degree have historically been used, and you are likely to come across various formats (Table 2-1). Degrees are conventionally divided into 60 minutes (also called minutes of arc, or arc-minutes), and each minute further sub-divided into 60 seconds (seconds of arc, or arc-seconds). These units are clearly not convenient for most digital applications, so the ‘decimal degrees’ format is usually used in GIS software.

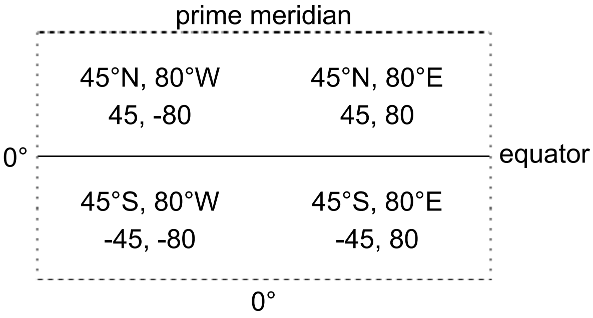
Table 2-1 Different units and formats for expressing latitudes and longitudes.

Degrees	D	52°N, 0°W
Decimal degrees	DD	51.521°N, 0.178°W
Degrees, decimal minutes	DDM	51°31’N, 0°10’W
Degrees, minutes, seconds	DMS	51°31’14”N, 0°10’39”W

2.1.2.2 Directions of angles: north – south, east – west

There is no standard for expressing the direction of angles. Printed references normally use ‘N’/‘S’ and ‘E’/‘W’, along with conventional signs for degrees (°), minutes (’) and seconds (”). When co-ordinates are stored digitally, however, these signs are not normally used, and directions are indicated by positive or negative numbers. Table 2-2 shows examples.

Table 2-2 Ways of expressing a co-ordinate in each of the four quadrants, showing a conventional printed format above and a digital format below.



45°N, 80°W 45, -80	45°N, 80°E 45, 80
45°S, 80°W -45, -80	45°S, 80°E -45, 80

When adding co-ordinates to your GIS from sources such as atlas, gazetteers or printed references, it is often necessary to convert from ‘DMS’ to ‘DD’. This is easily done in a spreadsheet, using a formula such as: DD = degrees + (minutes / 60) + (seconds / 3600). Care is needed to ensure that the right sign – positive or negative – is used.

2.1.2.3 Order: ‘latitude, longitude’ or ‘longitude, latitude’

A final difference is in the order. Co-ordinates are conventionally given with their latitude first, followed by their longitude: the shorthand “lat-long” is often heard. By contrast, most GIS software assumes, if not told otherwise, that co-ordinates follow the ‘x, y’ convention. In other words, longitude first, then latitude. If you are storing co-ordinates in a spreadsheet, make sure that columns are labelled clearly and if necessary change the order of the columns. I have overlooked this point more than once, only to find that a set of co-ordinates imported from a spreadsheet is mapped out ‘on its side’ in a GIS. This is particularly liable to happen when using geographic co-ordinates downloaded from a GPS (see Section 4.7).

What does a degree look like?

Having an idea of the ground distance represented by a degree helps to visualise latitudes and longitudes. The original definition of the metre is one ten-millionth of the distance between the equator and the poles: 90° of latitude = 10,000,000 m. Hence 1° of latitude = 111,111 m, or about 111 km. This is a useful rule of thumb when estimating the *precision* of a co-ordinate: the first decimal place corresponds to about 11 km, the second decimal place to 1.1 km, and so on. So, without doing any complicated calculations, we know that the latitude figure “51.52°N” is precise to about 1 km. Clearly, this applies only to degrees of *latitude*, which are always equally spaced (hence the name ‘parallel’ for lines of latitude). Degrees of *longitude* are also about 111 km at the equator, but from there become smaller as they converge towards the poles.

To summarise so far, we now have a reference system from which to measure positions on the earth’s surface. This gives us a precise way to specify the longitude and latitude of any feature, anywhere, on the earth, with reference to an ellipsoid; and we know how these positions relate to the actual shape of the earth – the geoid – by means of the datum.

2.1.3 From curved surface to flat map: map projections

Having established the locations of features on a rounded surface, the next step is to transfer them onto a flat surface – the process of map projection. Projections originated as a way of creating paper maps, but have an additional role in a GIS context: even if data are never plotted to a sheet of paper, much geographical analysis in a GIS relies on measurements of distances, lengths and areas. Such analysis is impractical or impossible using latitude/longitude co-ordinates, because they are angular measurements. Once these data have been projected onto a plane surface, the co-ordinates typically use metres or feet as units, so they become far more amenable to measurement and analysis, as well as being more intuitive and comprehensible to their human users.

The best known and 'original' projection is the Mercator projection (Figure 2-6), named after its inventor Gerardus Mercator, 1512-1594. Although it is well known now for its distortions in distances and areas, it has a unique application: a straight line drawn on a Mercator map corresponds to a compass bearing on the ground (or at sea), a property which was vital to the marine navigators who were (and still are) its main users. It is also worth noting that distances and areas are relatively accurate close to the equator, where the projection cylinder is in contact with the earth – this is known as the line of tangency.

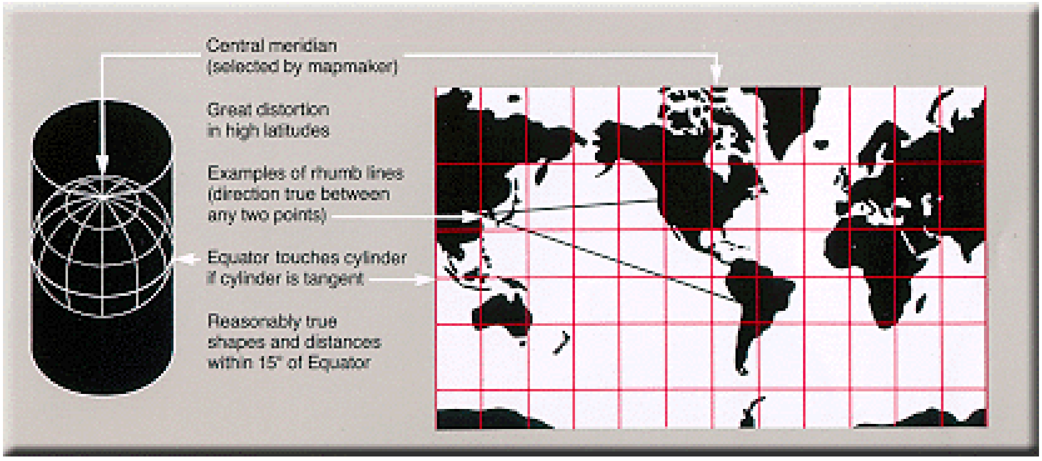

Central meridian (selected by mapmaker)
Great distortion in high latitudes
Examples of rhumb lines (direction true between any two points)
Equator touches cylinder if cylinder is tangent
Reasonably true shapes and distances within 15° of Equator

Figure 2-6 Construction of the Mercator projection (from http://erg.usgs.gov/isb/pubs/MapProjections/projections.html).

The example of the Mercator projection shows us that some compromise is inevitable when making a projection: while a globe can accurately portray directions, distances, areas and shapes, one or more of these properties becomes distorted when projected onto a plane. As a result, a huge variety of projections have been devised, each suitable for particular applications, different parts of the world and different areas of interest. Large scale topographic mapping at 1:50,000, for example, will use a different projection to a map showing population distribution over an entire continent.

Some projections commonly used for fieldwork and research are shown below, but for more details readers are referred to a US Geological Survey webpage: *erg.usgs.gov/isb/pubs/MapProjections/projections.html*

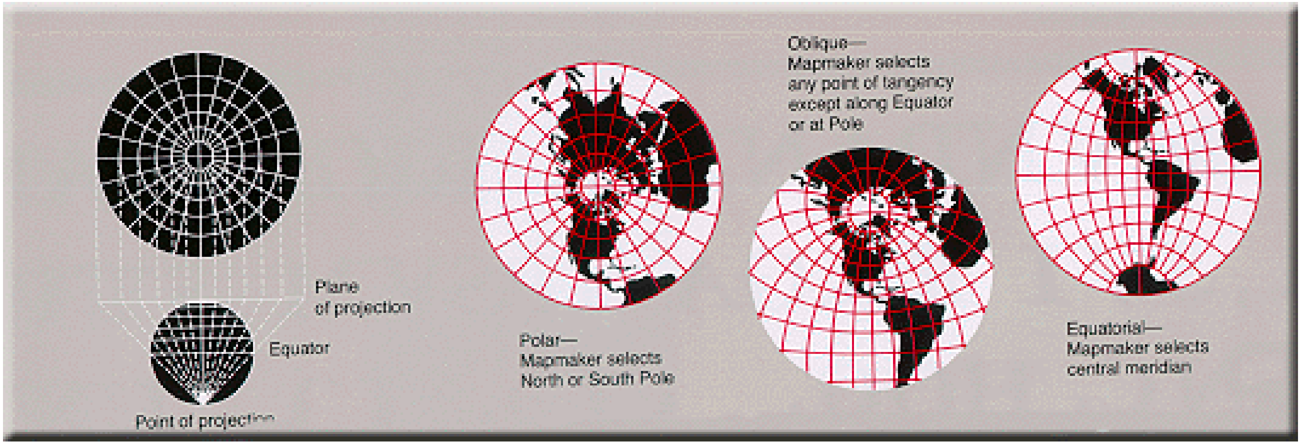


Figure 2-7 The stereographic projection, often used for mapping Arctic areas, and by the USGS and British Antarctic Survey for Antarctic mapping. Note that unlike the Mercator projection which uses a cylinder, it is projected onto a plane – this is known as an azimuthal *projection.*

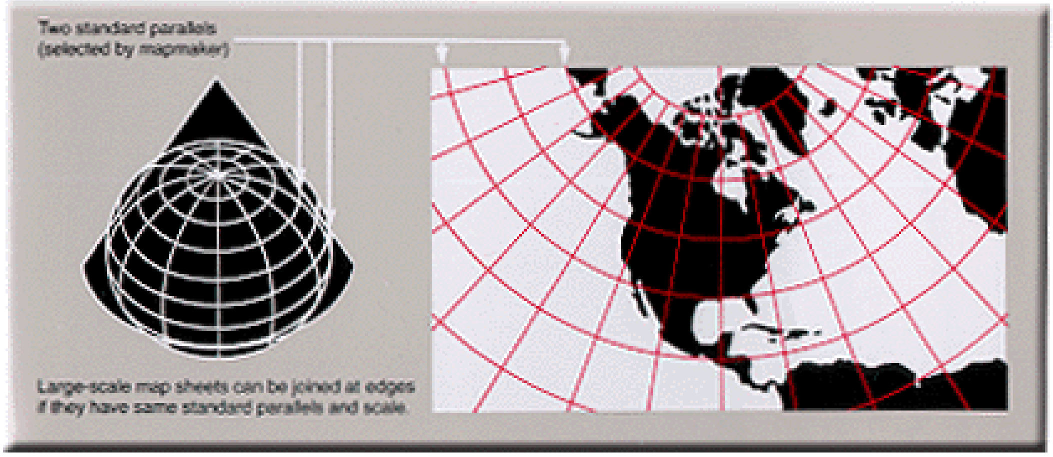

Two standard parallels
(selected by mapmaker)
Large-scale map sheets can be joined at edges
if they have same standard parallels and scale.

Figure 2-8 Lambert Conformal Conic projection, used for large-scale topographic mapping of the USA. Also suitable for topographic mapping of small areas or larger areas that run east-west. Following the previous two figures, this projection uses a third construction method, being projected onto a cone, hence the name conic.

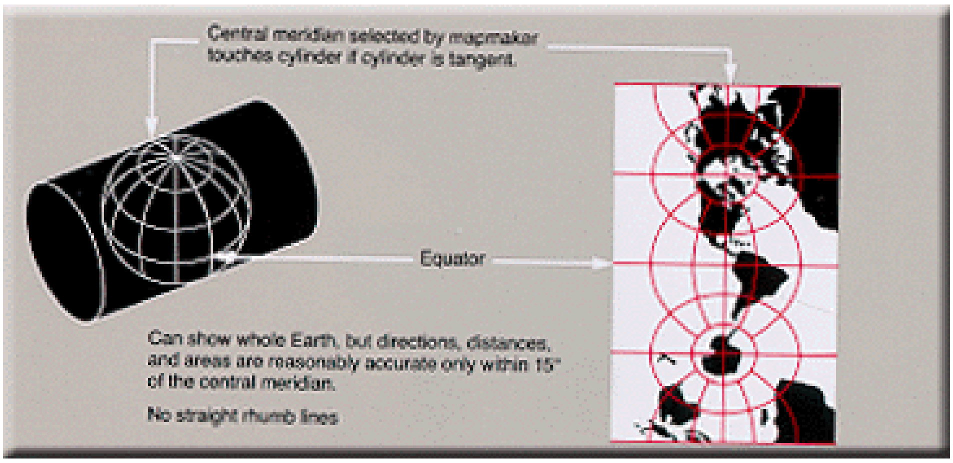

Central meridian selected by mapmaker
touches cylinder if cylinder is tangent.
Equator
Can show whole Earth, but directions, distances,
and areas are reasonably accurate only within 15°
of the central meridian.
No straight rhumb lines

Figure 2-9 Transverse Mercator: identical in construction to the Mercator, except that the cylinder is oriented transversely. The line of tangency, where scale distortion is at its minimum, therefore follows a meridian, i.e. a line of longitude; this property is used in the Universal Transverse Mercator co-ordinate system, described in a following section.

2.1.4 A rectangular co-ordinate grid

We now come to the final stage: once projected, a map is usually given a Cartesian co-ordinate system – one based on a flat, square grid – to refer to locations and to make measurements. This can be visualised as a grid overlaid on the map. Distances along the *x* axis are often referred to as *eastings* (i.e. measurements made in an eastward direction), while *y* axis distances are *northings* (measurements made northwards). The units are usually ‘real world’ ones: metres or (particularly in the USA) feet. An example of grid co-ordinates is shown in Figure 2-10, or look at any Ordnance Survey map to see similar examples.

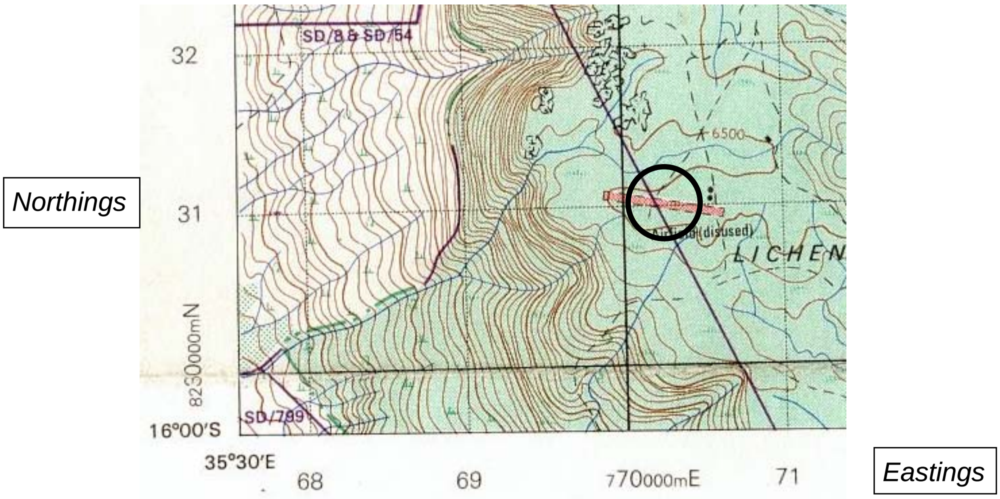


Figure 2-10 Scanned portion of a 1:50,000 map of Malawi showing its co-ordinate grid and the labels. Grid lines are 1,000 m (1 km) apart. Note how most of the labels comprise just two digits taken from the full co-ordinate. This is done to save space and to make map-reading easier. On this map, the complete co-ordinates are given every 10 km (in this case, 770000 east and 8230000 north). Note that the complete co-ordinate shows the unit (metres) and the direction (E/N). For example, the eastern end of the disused airfield (circled) is at: 770625mE, 8230950mN.

Most countries of the world have defined their own national grid systems, while an international system (UTM, detailed in the next section) is also widely used around the world. Expedition GISers therefore almost always use an existing national grid, for compatibility with local maps and with standard GPS/GIS settings. As an example, Section 11.4 shows the settings needed to make GPS readings correspond with co-ordinates on a 1960s map series of Vietnam. The following section now explains how such a co-ordinate grid is put together. There might be cases where a field project defines its own local grid system; this is unlikely, but might be done, for example, on a sub-Antarctic island, or where a local grid sampling system is needed. In such cases, if a grid is defined using the concepts shown here, it will be possible to transform from the local non-standard co-ordinate system to a standard system. A neat – if rather unusual – example was the creation of a new national grid for the Maldive islands; because they straddle the equator (just), UTM would have been inappropriate, so a modified UTM grid was defined, with an origin 100 km south of the equator (Hobbs 2003).

How is the co-ordinate grid positioned in relation to a map that has been projected? Horizontally, it is usually aligned with the central meridian (line of longitude) of the map projection. This means that the grid is upright – north-aligned – at the centre of the mapped area. Vertically, it is usually based on a particular parallel (line of latitude), often the equator, although other parallels may be used. The British national grid, for example, is aligned horizontally with the central meridian of the underlying map projection (2°W) and vertically with a parallel just to the south of the British land mass (49°N).

The grid also needs an origin, the point defining the 0,0 co-ordinate. The 'natural' origin for x co-ordinates (eastings) is the central meridian of the map projection. However, to avoid negative values occurring to the west of this line, a 'false easting' is typically used, as shown in Figure 2-11. This is a value added to all *x* co-ordinates, in effect creating an origin to the west of the area being mapped. Thus all *x* co-ordinates within the mapped area will be positive. In the British national grid, for example, the false easting is 400,000, which means that eastings are measured from a line lying offshore to the west of Britain.

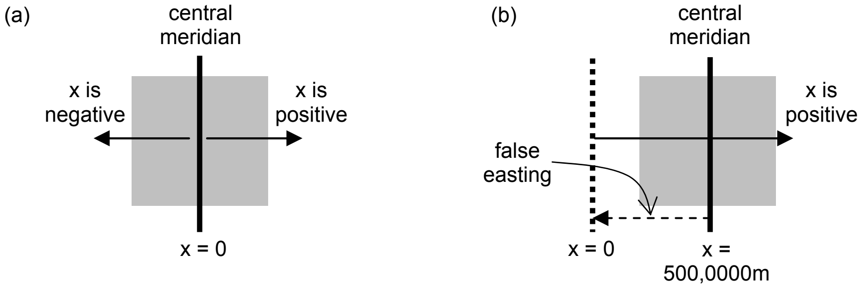


Figure 2-11 Using a false easting to create positive co-ordinates within the mapped area (marked in grey). (a) Using the central meridian as the origin results in negative values to the west. (b) Adding a false easting of 500,000m ensures positive values throughout the mapped area.

The origin for *y* co-ordinates (northings) is usually a particular parallel (line of latitude). In many cases the equator is used; this provides a 'natural' latitude of origin. Alternatively, a parallel closer to the area being mapped may be used: in the British example, it is 49ºN. Again, an offset – the false northing – may be applied to avoid negative numbers. The false northing value is added to all *y* co-ordinates to make them positive. Figure 2-12 shows a common case, where a large false northing is applied in order to make southern hemisphere co-ordinates positive when measured relative to the equator. The case of the British national grid is slightly unusual: negative northings do not occur in any case, as the latitude of origin has been chosen to lie south of the British landmass. However, a false northing of 100,000m is still used, in order to bring the origin closer to the mapped area.

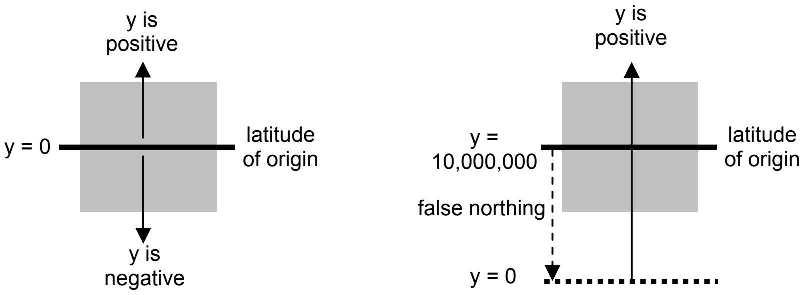
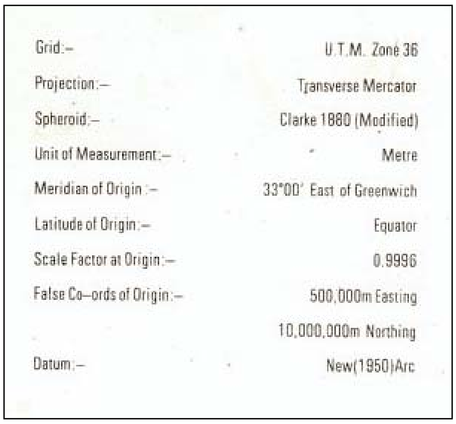


Figure 2-12 Using a false northing to create positive co-ordinates within an area being mapped (shown in grey). In this case, the false northing is 10,000,000m.

The South African grid system is an exception: it does not use a false northing, so co-ordinates have an increasingly large negative value as one moves southwards from the equator. This system is shared by Lesotho and Swaziland (Chief Directorate: Surveys & Mapping 2003).

We now have all of the information necessary to describe how the position of a point – the mountain we are climbing or the tree we are studying – can be (1) measured (using a datum), (2) mapped (using a projection) and (3) given a co-ordinate (using a grid). At best, maps show all of these details (as in Figure 2-13), allowing you to configure your GPS and GIS correctly. Similarly, all GIS data should have this information specified, either within a particular file format or in an associated document (metadata). Often, however, some further detective work may be required to find all of the parameters, for example by asking a national mapping agency.



Grid:–	U.T.M. Zone 36
Projection:–	Transverse Mercator
Spheroid:–	Clarke 1880 (Modified)
Unit of Measurement:–	Metre
Meridian of Origin :–	33°00' East of Greenwich
Latitude of Origin:–	Equator
Scale Factor at Origin:–	0.9996
False Co–ords of Origin:–	500,000m Easting
	10,000,000m Northing
Datum:–	New(1950)Arc

Figure 2-13 Full details of the projection and co-ordinate system printed in the margin of the Malawi map shown in Figure 2-10.

2.2 The Universal Transverse Mercator (UTM) system

UTM is not a particular projection or co-ordinate system, but rather defines a set of map projections and co-ordinate systems designed for large scale mapping in all parts of the world. Many national mapping agencies use UTM for their topographic map series; Landsat images provided free by the Global Land Cover Facility (University of Maryland; see Chapter 5) use UTM; and it is supported by almost all GIS software and GPS receivers. Thus it is a common choice for expeditions, whether researchers, adventurers, or both.

2.2.1 Map projection and UTM zones

As the name suggests, UTM uses the Transverse Mercator projection. In fact, it comprises 60 different Transverse Mercator projections, each one with a central meridian 6º greater than the previous one. Visualise the cylinder in Figure 2-9 being rotated in 6º increments, each time forming a new central meridian where it is tangent with (touches) the earth's surface. This results in 60 'zones' around the world, each 6º wide, within which map distortions are insignificant. In the northern hemisphere these zones are numbered 1N to 60N, while in the south they are 1S to 60S (Figure 2-14). Thus, for example, most of Madagascar lies in Zone 38S, which spans from 42ºE to 48ºE, with a central meridian of 45ºE.

2.2.2 UTM co-ordinates

The unit for all UTM co-ordinates is metres. Northings (*y* co-ordinates) are measured with reference to the equator, increasing northwards from 0 m at the equator. In the southern hemisphere, a false northing of 10,000,000 m is applied, to ensure positive co-ordinate values (see Section 2.1.4).

Note that all *y* co-ordinates, whether north or south of the equator, therefore lie within the same range of values (0 to 10 million): a *y* value by itself gives no indication of which hemisphere it is in. It is therefore vital to specify whether a co-ordinates is in a southern or a northern UTM zone.

UTM eastings (*x* co-ordinates) are measured with reference to the central meridian of the zone, with a false easting of 500,000 m. In this case, note that there is no indication in the co-ordinate itself of its zone; a given easting could exist within any one of the 60 zones, so be sure to specify the zone number.

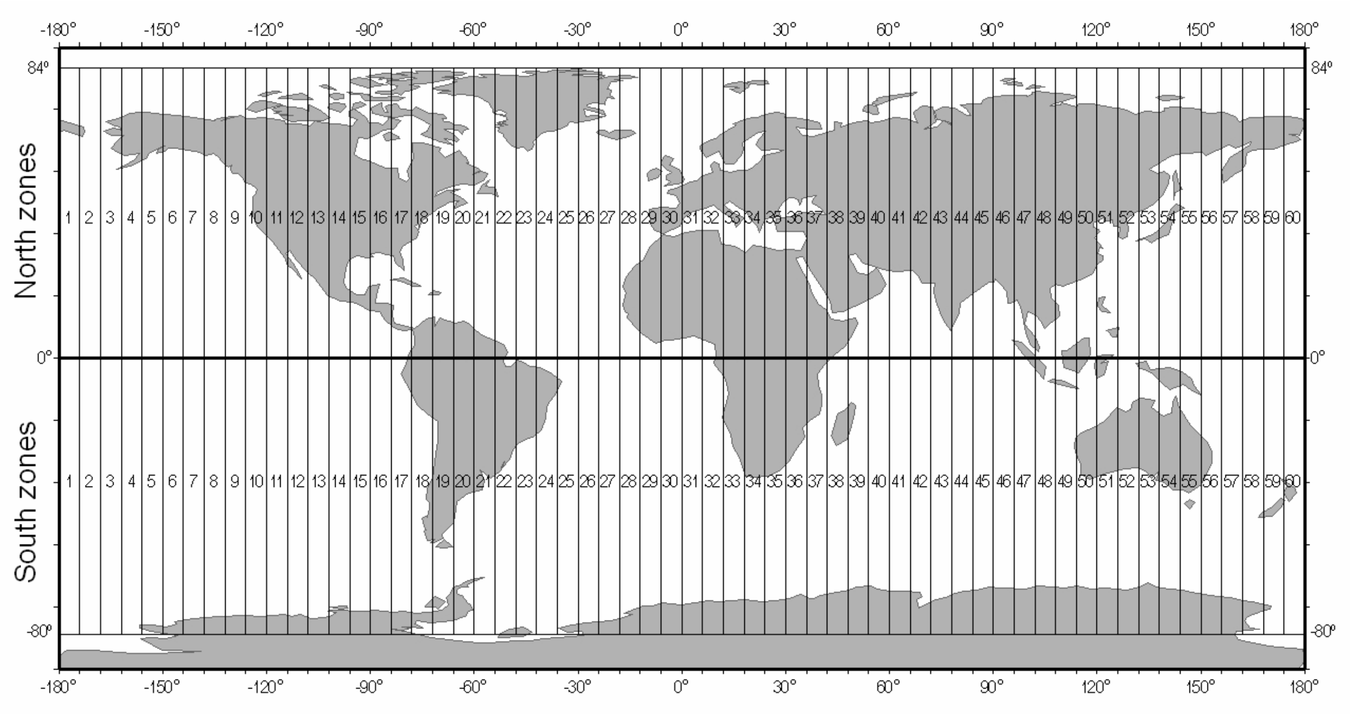


Figure 2-14 The UTM zones, numbered 1-60. Northern ones are denoted as Zone 1N, 2N, etc, and southern ones as 1S, 2S, etc. Most of Madagascar, for example, is in Zone 38S.

2.2.3 Datum

While UTM specifies a projection and grid co-ordinate system, it does not specify a particular datum, leaving the choice of datum to be appropriate to the local mapping area. Indeed, the UTM definition has existed since the 1940s, well before global geocentric datums such as WGS84 came into existence. Figure 2-13 for example shows the details of a UTM map that uses the Arc 1950 datum. More recent UTM maps and GIS data sets, however, tend to use WGS84, and this is the choice of the Landsat images available on the GCLF website (http://glcf.umiacs.umd.edu/data/).

2.2.4 Scale factor in UTM

Within each zone, distortion in scale is minimised by the use of a scale factor. Normally, the scale of a map is 'true' along the central meridian of a Transverse Mercator projection; the scale factor is said to be 1. To the east and west, however, scale is increasingly distorted: the scale factor increases above 1, and the effect is to increase distances between projected points.

While this effect is negligible close to the central meridian, the distortion increases steeply away from centre towards the edges of the UTM zone. To minimise this effect, the east-west scale across the zone is reduced by a given scale factor, applied from the central meridian outward. In the UTM system (as with British OS maps), the scale factor is 0.9996. This means that the scale along the central meridian is reduced by this factor, but the advantage is that there are then two lines, one to either side of the central meridian, along which the scale factor is 1 – no scale distortion. Further out towards the edges of the zone, scale distortion is significantly reduced. Figure 2-13 shows the scale factor quoted on the Malawi map example.

What happens at the edge of a UTM zone

Reducing scale distortion at the edges of UTM zones has a useful effect for fieldworkers: if your study area happens to stray across the defined limit of a UTM zone (we cannot expect national parks or species ranges to stay within UTM zones!), then there is little harm in keeping to the same UTM parameters beyond the zone limits. Figure 2-16 shows the distortion that occurs when the same projection is used to map areas far outside a zone: distortion is large far away from the central zone, but immediately to each side the distortion is, for most purposes, insignificant.

Indeed, some single-sheet national maps use precisely this method. Although Tanzania spans three UTM zones (35S, 36S, 37S), Harms-Verlag based its recent 1:1,400,000 map of Tanzania on the projection for the UTM zone at the centre of the map, i.e. zone 36S. Note that at this scale, distortions are not significant, but using the same UTM system across such a wide area would not be acceptable on 1:50,000 maps, as distortions would become far more apparent. However, it does demonstrate that this method can be applied, with care, in appropriate situations.

This method will work when storing data and creating maps in a GIS. However, GPS units will always display the 'right' UTM zone, calculated on the basis of their known longitude; they cannot normally be forced to display co-ordinates for a UTM zone that they are not in.

2.2.5 Example of a UTM co-ordinate

The following two figures show how the location of a mountain peak in southern Ethiopia is mapped to a UTM co-ordinate. In this case, the full co-ordinate is:

700600mE, 987700mN, UTM zone 36N, WGS84 datum.

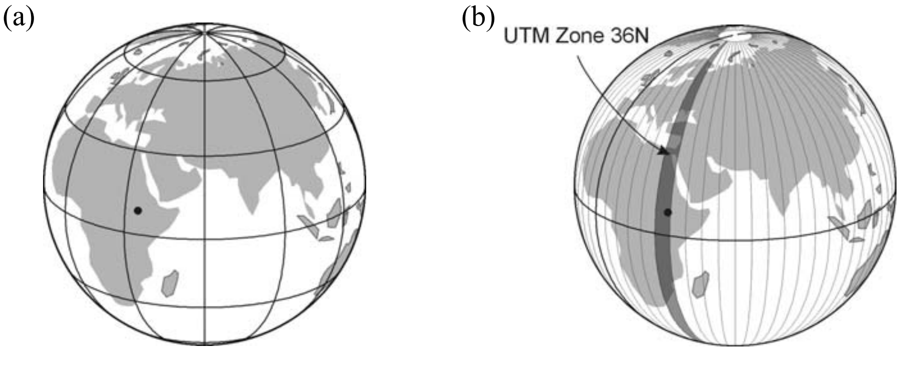


Figure 2-15 (a) Location of the point on the globe. (b) The relevant UTM zone is 36N.

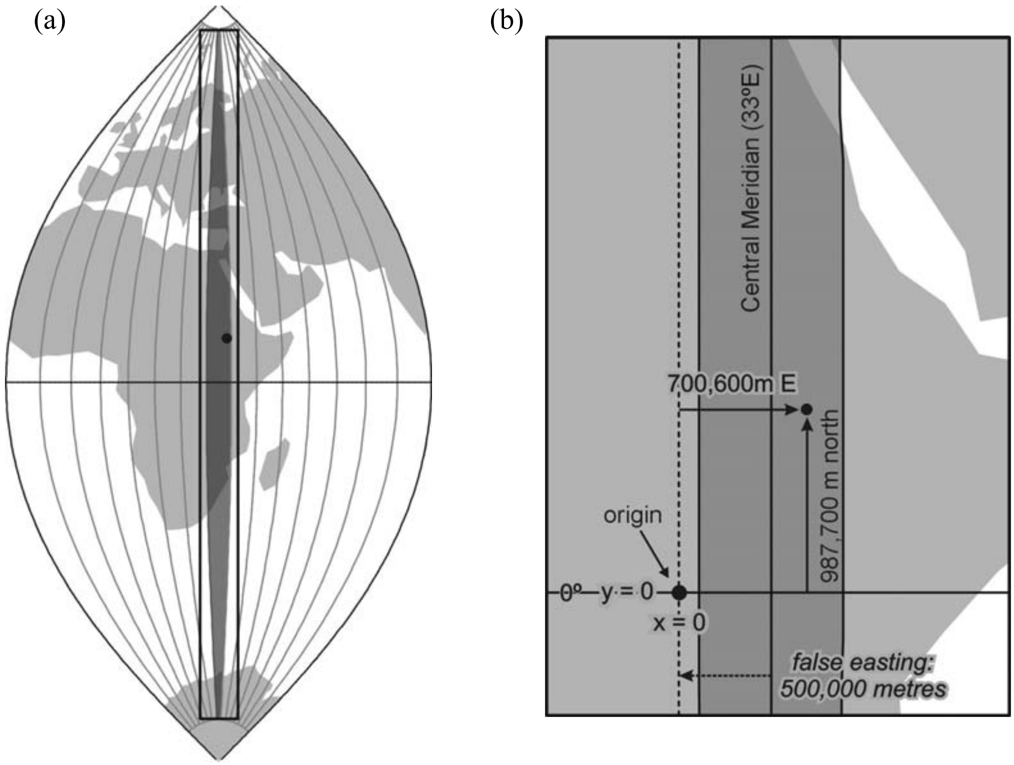
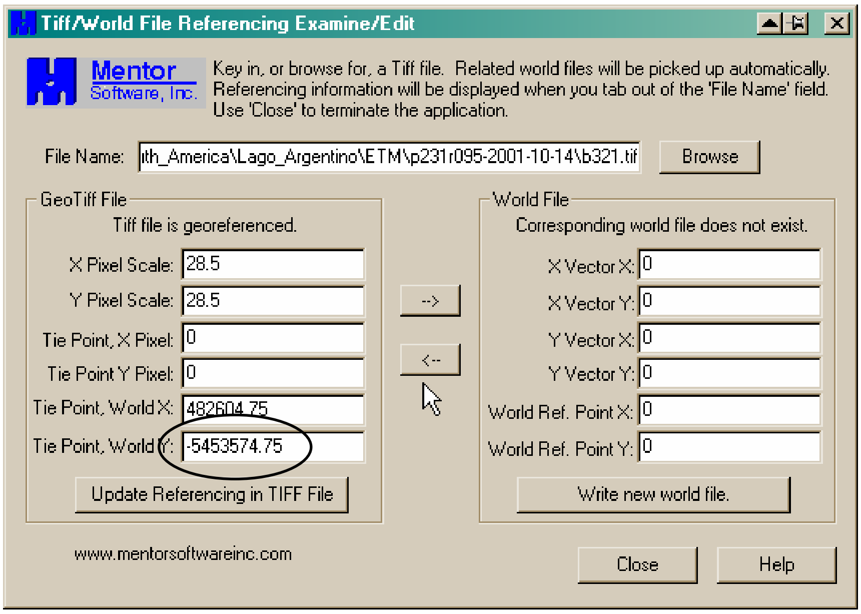


Figure 2-16 (a) UTM zone 36 after being projected onto a plane using Transverse Mercator. Adjacent UTM zones are also displayed to show the effect of scale distortion away from the central meridian. The black box shows the rectangular extent of the co-ordinate grid. (b) Detail showing the x *co-ordinate in relation to the central meridian, with a false easting of 500,000m, and the* y *co-ordinate in relation to the equator. Being in the northern hemisphere, there is no false northing.*

Landsat images from the Global Land Cover Facility website: which UTM zone?

There appears to be a UTM zone problem with some of the free satellite images available from the GLCF website, affecting Landsat ETM images in the southern hemisphere that use the TIFF file format. The co-ordinates of these images are given as negative values; in other words, no false northing has been applied, and the images are incorrectly referenced to UTM north zones.

This can be corrected with a free utility programme called GeoTiffExaminer, which can read and modify the geo-referencing information held in GeoTIFF files. GeoTIFFs are the same as normal TIFF bitmap files, but with added information about their geographical co-ordinates contained in their headers. Use GeoTiffExaminer to read the header information of a TIFF file (Figure 2-17), and note the negative value of the 'Tie Point, World Y'. Add the false northing (10,000,000) to this number and click 'Update Referencing in TIFF File'. This corrects the co-ordinate referencing error. Then when you open the image in a GIS programme, you also need to modify the UTM zone designation. For example, if the zone is read by the software as 18N (or if it is not read at all), then change it to 18S. The image should now be correctly referenced, and other data layers should overlay it in the right positions. This is a fine example of a problem whose solution requires an understanding of co-ordinate systems!



Tiff/World File Referencing Examine/Edit

Mentor Software, Inc.

Key in, or browse for, a Tiff file. Related world files will be picked up automatically. Referencing information will be displayed when you tab out of the 'File Name' field. Use 'Close' to terminate the application.

File Name: ith_America\Lago_Argentino\ETM\p231r095-2001-10-14\b321.tif

GeoTiff File

Tiff file is georeferenced.

X Pixel Scale: 28.5

Y Pixel Scale: 28.5

Tie Point, X Pixel: 0

Tie Point Y Pixel: 0

Tie Point, World X: 482604.75

Tie Point, World Y: -5453574.75

World File

Corresponding world file does not exist.

X Vector X: 0

X Vector Y: 0

Y Vector X: 0

Y Vector Y: 0

World Ref. Point X: 0

World Ref. Point Y: 0

Figure 2-17 Using GeoTiffExaminer to correct an error that occurs in some southern hemisphere Landsat ETM images. Increase the circled value by 10,000,000, in this case to 4,546,425.25. Then in a GIS programme, specify a southern rather than northern UTM zone.

2.3 Choosing a co-ordinate system for your GIS

2.3.1 Are you working with an existing GIS?

If someone else has already established a GIS of your study area, it is worth considering the same co-ordinate system as theirs: you are able to use their data to get started, and later on will be able to share your data with them. Check that your GIS software and your GPS unit support the system.

Example: in 2002 a small team helped map the road network and establish a field GIS in Mkomazi Game Reserve, Tanzania. They based the GIS on the same co-ordinate system as the one used by an RGS project in Mkomazi during the 1990s. The team was thus able to use existing GIS data layers, which were available for download on the internet. In this case, the system was: UTM, zone 37 south, Arc 1960 datum.

2.3.2 Will you be working with existing maps?

If detailed local maps already exist, for example at a scale of 1:50,000 or 1:100,000, then using the same co-ordinate system as the maps has several advantages: (1) a suitable projection and datum has been chosen for the area; (2) features can be digitised from the maps without the need for re-projection; (3) and you will be able to read the maps using the same co-ordinate system as your GIS, making for easier planning and navigation.

Example: the same team was asked to produce a road map of Mikumi National Park in Tanzania, for use by both researchers and tourists. 1:50,000 topographic maps were available, so they based the GIS on the same co-ordinate system as the maps (UTM / Arc 1960 datum). GPS units were also set to the same system. Thus park staff and researchers are now using the maps, GPS and GIS, all with the same co-ordinate system and without need for conversions.

2.3.3 Do you have existing data already geo-rectified and geo-registered?

It might be that one existing data set will determine the co-ordinate system you use. If you start with a series of large raster images, for example, it would be time-consuming to re-project them, and you would loose some accuracy in the process. The issue of accuracy is particularly important if you are undertaking any numerical analysis of the raster data sets: the actual values of each pixel are not very important if the image is just being displayed or printed as a map, but if it is being used for change analysis or correlation, then pixel values are critical. If the images are already in an appropriate projection for the study area, it might be best to keep that projection, and re-project other vector data to match it. In most cases, vector data can be re-projected with little or no degradation.

Example: several sets of environmental data were used for a GIS model of mammal distribution in Mkomazi Game Reserve, Tanzania. The main set comprised satellite imagery (AVHRR), downloaded from the internet. This comprised images taken every 10 days over two years, each with several different wavebands, making hundreds of images in all. They were already geo-registered and projected in Sinusoidal projection. All my other data, however, used the same projection as the Tanzanian 1:50,000 maps. It was far easier to re-project the vector data to the Sinusoidal projection than to re-project hundreds of satellite images.

2.3.4 Starting from scratch?

If there is no existing mapping or GIS data, what system to use? If you find yourself in the field with no idea which co-ordinate system to use, the best option is to save data in unprojected co-ordinates – in other words, latitude and longitude – using WGS84 datum. This provides the maximum flexibility for transforming the data to other systems should different needs arise in the future.

Where a paper map is to be printed or analysis to be undertaken, then UTM typically provides a good answer, being designed for large scale use anywhere in the world. It is also commonly recognised and accurately handled by most GIS software and GPS units.

Other considerations may apply to more specialist applications, particularly in large study areas at country or continental scale, although these tend to be outside the scope of most expeditions.

Example: Bogda Shan expedition in NW China was mapping a relatively small mountainous area with no existing detailed maps. They chose the UTM co-ordinate system, with the WGS84 datum.

2.4 Conclusion

Perhaps most importantly is to (i) know which projection, datum and co-ordinate system you are using and (ii) always record it, in fieldnotes, the GIS, and reports and publications. Why? You can re-project the data if you decide another system would be better; and others will be able to use your valuable geographical information.