30

Spatial referencing and coordinate systems

H SEEGER

The growing use of the Global Positioning System (GPS) offers an apparent panacea for many GIS-related problems. However, this is partly illusory: to avoid subsequent errors when matching GPS coordinates to those gathered from pre-existing sources – a normal requirement – a good understanding of geodesy is essential. This chapter describes the basic principles of geodetic reference systems, geodetic ellipsoids, and map projections. It shows the nature of the relationship between WGS84, which is based on geocentric Cartesian coordinates, and national systems based on an ellipsoid defined locally to suit the national territory. Most geographical data have hitherto been collected on maps which are themselves based on nationally defined reference systems and a variety of other factors has entered into their construction and digital encoding. As a result, coordinate shifts as large as hundreds of metres may occur in different descriptions of the same place. The ramifications for the use of GIS to link data together are obvious.

1 INTRODUCTION

Only in a few, very local, applications of GIS can geodetic reference systems and coordinate systems be ignored safely. Such situations arise when a particular study relates only to an area a few tens or hundreds of metres across and involves no comparisons with studies elsewhere. In most other circumstances, however, it is essential that the geographical descriptions are made using a more global reference framework (see also Dowman, Chapter 31; Guptill, Chapter 49; Smith and Rhind, Chapter 47). Typically in the past this has been achieved through use of national geodetic frameworks defined and maintained by national mapping agencies and manifest most obviously on topographic maps. Now the advent of the Global Positioning System (GPS: see Lange and Gilbert, Chapter 33) has offered the possibility of a uniform reference framework for the whole world and a means of determining position within that framework at low cost and instantaneously. It is self-evident therefore that these developments have major implications for GIS.

In practice, these developments have even greater ramifications for GIS users because the apparent de-skilling of surveying and the globalisation of reference frameworks conceals a number of major complexities. It is entirely possible that serious errors can be made with the new tools, especially when their use is (necessarily) combined with information derived from earlier surveying concepts and methods. In Bosnia, for instance, peace-keeping troops using GPS to navigate by jeep have been killed by mines because of disjunctions between datums underlying the maps of minefields and the GPS itself. Serious naval incidents have occurred because of the same factors and blind trust in the new technology. It follows that serious use of GPS – and indeed of GIS – requires a basic understanding of the different geodetic frameworks and the relationships between them.

This chapter examines the basic concepts underlying methods of defining location and the description of that positional information. There is not space enough to provide comprehensive guidance on, or the mathematical basis of, a complex topic. Nor can it give recommendations on which particular system to use in specific circumstances. It does, however, give sufficient information to enable the reader to understand some of the benefits and pitfalls associated with choosing or coping with different reference systems. Standard texts on geodesy, such as that by Bomford (1980) and Vanicek and Krakovsky (1986), provide detail on background concepts while Lange and Gilbert (Chapter 33) and books by Kennedy (1996), Leick (1990), and Hofmann-Wellenhof et al (1994) give more details on GPS.

2 SOME BASICS ABOUT GEODESY

Geodesy is the science of the shape and size of the Earth, together with its gravity field. It is well-known that the Earth is not a regular geometrically-shaped body, and that variations in its shape occur continuously at the meso and micro scales. In many instances (e.g. in Alpine, Himalayan, or Andean areas), it is simply impossible to make good local approximations by regarding the Earth's level surfaces as flat planes.

In mathematics, there are in general two ways to determine a point position in space – that is, by Cartesian coordinates or by angular coordinates. The coordinates of the point are normally related to a right-handed rectangular coordinate system, or other combinations. The position of the origin of the coordinate system and its orientation is given by the geodetic datum definition.

2.1 Early measurements of the Earth

Eratosthenes was one of the first to measure the size of the Earth with any accuracy but provably accurate definitions of the size of the planet only came about as a result of work in the eighteenth century. French expeditions in 1735 to both Lapland and Peru measured along arcs of meridians using triangulation. The measurements proved that the Earth was almost spherical but flattened at the poles (oblate) and not at the equator (prolate), thus being consistent with Newton's gravitational theories.
As a result of triangulation and astronomical observations near Dunkirk and Barcelona, Jean-Baptiste-Joseph Delambre (1749–1822) and Pierre-François-Andre Mechain (1744–1804) proposed the definition of the metre as one ten millionth of a quadrant of the meridian through the Observatory of Paris. This definition was accepted by the French Academy of Sciences in 1795. The use of such universal standards of time and length are crucial to geodesy.

2.2 Geoid and ellipsoid

Defining a single, stable reference surface from which the third dimension is measured was also found to be difficult. The most obvious choice of reference surface is mean sea level but this is not the same everywhere: it is, for instance, affected by ocean currents, wind, and barometric pressure. As a result, a conceptual surface called the *geoid* is used. The geoid is the equipotential surface of the Earth's gravity field that best approximates to mean sea level. An equipotential surface is level in the sense that moving across it results in no work being done against gravity (i.e. the direction of gravity is everywhere perpendicular to it). A simple conceptualisation of the geoid is to imagine the Earth with canals dug across all continents, enabling completely stationary and inert oceans to meet and find their own 'sea level' assuming there were neither tides or weather to affect the result. Even now the geoid is difficult to model: world-wide it is known to a metre or so, although locally it can be found to within 0.01 metres.

The geoid is a smooth surface which can be approximated by an *ellipsoid*. An ellipsoid is defined as an ellipse of rotation about the polar axis. The relationship between these two shapes – the geoid defined by physics and the ellipsoid defined by mathematics – is of prime importance, as will be evident later. This relationship is part of what is termed the geodetic datum (see below).

A further complication is that the Earth is not a rigid and invariant body. Sea tides are quite obvious to observers at the seashore but the same forces also cause the solid Earth to have tides, albeit much smaller ones of only a few centimetres. Because the Earth is elastic, has an atmosphere, is rotating in space, and orbiting the Sun, the poles move. The poles move (polar motion) with both long- and short-term periods. Fortunately, as the instrumentation and methods available to measure the Earth improve, all these temporal movements become measurable but they are time-dependent and thus need to be repeated periodically. They also need to be updated as instrumentation and methods develop.

2.3 Contemporary approaches

From the mid-twentieth century onwards, measurement techniques using light waves and microwaves were considerably improved and the ability to measure long distances (70–100 kilometres) was used by geodesists to refine the nineteenth-century triangulation schemes and re-define the shape of the Earth. The latest methods of obtaining the Earth's size and gravitational field all involve looking outward. Very long baseline interferometry (VLBI) relies on two astronomical telescopes at each end of a baseline (e.g. in Santiago, Chile and in Wetzell, Germany) correlating the signals received from the same quasar. The time delay between one wave-front arriving at the telescope and the same wave-front arriving at the other telescope is measured with a very precise clock. Since the speed of propagation of the wave is known, the distance and direction can be calculated between the two telescopes. The accuracy of these baseline measurements is about one part in 100 million.

In addition, satellites are now used increasingly to determine baselines (distances and directions) and gravitational forces. The most widely used systems come from the USA. The first was TRANSIT, a US Navy navigation satellite which enabled geodesists to measure relative positions over hundreds of kilometres to an accuracy of around 30 centimetres using the Doppler effect of the satellite signals. The second was the GPS which has a constellation of over 24 satellites orbiting the Earth and enables positions between receivers to be deduced to one part in 1000 million using the signal phase differences (see Lange and Gilbert, Chapter 33). Russia has launched a similar system called GLONASS developed initially, like GPS, for military navigation. Satellite laser ranging (SLR) to orbiting satellites from ground stations (e.g. at Herstmonceux, England and Zimmerwald, Switzerland) enables the orbits of such satellites as LAGEOS to be determined and subsequently the effect of the gravity field to be computed. Satellite altimetry using radar to measure from satellites (e.g. ERS-1, ERS-2, and TOPEX POSEIDON) to the ocean surface allows the geoid to be modelled. The latest geodetic methods are also used to determine tectonic plate motion and for earthquake prediction (see, for example, Warita and Nonomura 1997).

3 THE DEFINITION OF POSITION

In this section, a more formal approach is used to define the way in which geographical location or position may be described. Two main types of position information can be distinguished:

- direct positioning – that is, based on coordinates which enable any point in space (one, two, three, or even more dimensions) to be uniquely defined and referred to a (local) coordinate system;
- indirect positioning – that is, based not on coordinates but on values (for instance administrative units, postal addresses, and road numbers) which can be unambiguously mapped – to some defined level of precision and accuracy – to a specific geographical location.

3.1 Direct position

'Direct position' is given by a set of coordinates based on a reference system in which all the physical parameters of the Earth – such as its size, shape, orientation in space, and gravity field – are defined, together with a specific coordinate system. It was pointed out earlier that the Earth itself changes in shape but, in addition and because our knowledge of the Earth is increasing, the values which we gave to some physical constants some years ago are not the 'best' values today. The effect of all this is that reference systems for the Earth, which by their very nature include the physical constants, change periodically.

There are consequently two aspects of direct positioning to be taken into account. The first and essential aspect is that a full coordinate description is required of point, line, area, or volumetric entities. The second aspect is often optional (but worthwhile): it consists of all necessary data to transform these coordinates into coordinates in another reference system.

3.1.1 Geodetic reference systems

All direct positions are defined by a geodetic reference system, including a *geodetic datum*. This contains all the elements necessary to locate a point relative to the surface of the Earth. A geodetic reference system is uniquely identified by an attribute name which is a short acronym or an abbreviation of the full name (e.g. WGS84 for the World Geodetic System 1984). The full name will usually contain the complete name of the coordinate reference system. The optional description attribute contains a more detailed textual description of the coordinate reference system, including recommended usage and possible pitfalls.

The origin, orientation, and rotation of the coordinate system is defined by the datum definition attribute (see section 3.1.4). Most geodetic reference systems have only one datum. However, because of the historic difficulty of surveying and computing horizontal position (bi-dimensional) and vertical position or height (uni-dimensional), the horizontal and vertical components are often separated. Thus a coordinate reference system may contain two datums, the geodetic datum and the vertical datum. However, hybrid datums are here solved as a combination of a bi-dimensional and uni-dimensional geodetic datum within the same datum definition. An optional list of 'anchor', or fiducial, points gives additional positional information which completely defines geodetic reference system. In summary, then, the coordinate system describes the spatial parameters of the 'measurement framework' and how it relates to the Earth as a whole; while coordinates define a 'direct position' within that system.

The easiest way to define a point in space is to use 3-dimensional Cartesian coordinates. Such a geodetic reference system has:

- an origin, *O*, generally coincident with the centre of mass of the Earth;
- one axis, *Z*, coincident with the Earth's spin axis;
- one axis, *X*, lying in the plane of the zero meridian (now close to, rather than exactly at, Greenwich);
- the *Y* axis completing the mutually perpendicular axes, forming a right-handed system.

For any point *P* in space, the choice of a geodetic reference system enables Cartesian coordinates *X*, *Y*, and *Z* to be defined as shown in Figure 1.

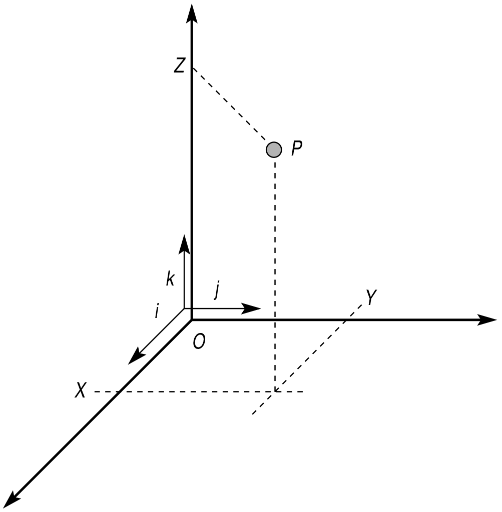


Fig 1. Cartesian coordinates on a geodetic reference system.

3.1.2 Geodetic ellipsoids and the zero meridian

As described earlier, geodetic studies in the eighteenth century showed that the mathematical figure which best represents the Earth surface without topography is the geoid and that this can be represented by an ellipsoid of rotation. This is termed a *geodetic ellipsoid*. It is defined by two parameters, for instance the semi-major axis and the 'flattening'.

The simultaneous selection of a geodetic reference system (which defines the 3-dimensional framework) and a geodetic ellipsoid (which defines the geodetic reference surface) is the basic model which distinguishes between horizontal and vertical information through (geodetic) geographical coordinates. When this choice is made, it is assumed that the origin of the ellipsoid coincides with the origin *O* of the reference system and that the minor axis (of rotation) of the ellipsoid coincides with the polar axis of the geodetic reference system. We can then define the geographical coordinates Latitude, Longitude, and Height (φ, λ, h) of any point, *P*, in space (as shown in Figure 2).

In Figure 2 *p* is the projection of *P* on the geodetic ellipsoid along the normal to the ellipsoid at *P*, and *h*, the ellipsoidal height, is the algebraic

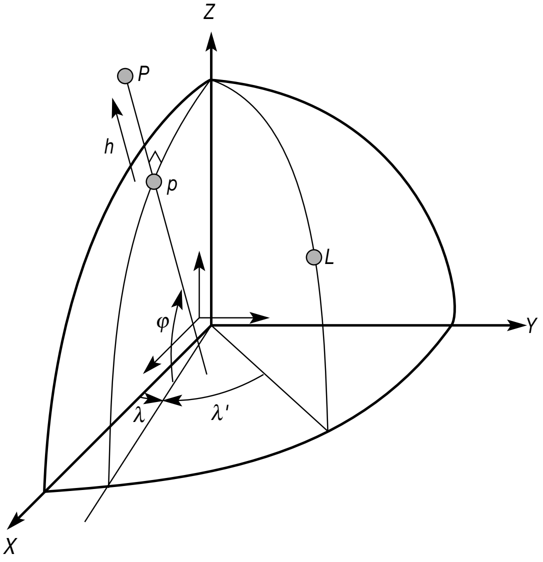


Fig 2. Geographical coordinates.

length of the segment (*pP*) (which is positive outside the ellipsoid). The latitude φ is the angle between the equator of the ellipsoid and the normal to the ellipsoid at *P*. The longitude λ is the angle directed between the Greenwich Meridian plane and the meridian plane of *P*. If another zero meridian is chosen (such as the meridian plane passing through a given point *L*), a different longitude λ' naturally results (as in some French maps before universal adoption of the Greenwich meridian). Following from this, a geodetic reference system, a geodetic ellipsoid, and a zero meridian allow a second type of 3-dimensional coordinates, the *geographical coordinates* (φ, λ, h), to be defined for any point *P* in space.

The scientific community has adopted a conventional terrestrial reference system (CTS), which is fixed to the Earth, to be used as reference for scientific and practical works. Its detailed definition and practical implementation have been improved continually. These tasks were the responsibility of the Bureau International de l'Heure (BIH) until 1987 and have been carried out by the International Earth Rotation Service (IERS) since 1988. This system, now named the IERS Terrestrial Reference Frame (ITRF), is truly geocentric and its orientation is the reference orientation of the planet for all Earth rotation studies.

The following relation exists between the ellipsoidal and Cartesian coordinates if the origin of the Cartesian coordinate system coincides with the centre of the rotation ellipsoid, the *X*-axis of the ellipsoid pierces the ellipsoid at the point $\varphi = 0$, $\lambda = 0$ and the *Z*-axis stands perpendicular to the equator:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} [N_E + h]\cos\varphi \cos\lambda \\ [N_E + h]\cos\varphi \sin\lambda \\ [N_E(1-e^2)+h] \sin\varphi \end{bmatrix}$$

With the transverse curvature radius

$$N_E = a\,(1-e^2 \sin^2 \varphi)^{-\text{q-}\Sigma}$$

and the first eccentricity

$$e^2 = 1 - \frac{b^2}{a^2}$$

we obtain

$$\lambda = \arctan \frac{y}{x}$$

The inverse problem of φ and h can be solved only by iteration, but the solution converges quickly since $h << N_E$.

$$h_i = \frac{(x^2 + y^2)^{\text{q-}\Sigma}}{\cos\varphi_{i-1}} - N_{E(i)}$$

$$\varphi_i = \arctan \left[\frac{z}{(x^2 + y^2)^{-\text{q-}\Sigma}} - \frac{1}{1 - \frac{e^2 \cdot N_{E(i)}}{N_{E(i)} + h_i}} \right]$$

The iteration may be started with

$$\varphi_0 \approx \arctan\left(z \cdot (x^2 + y^2)^{-\text{q-}\Sigma} \right)$$

3.1.3 Coordinate systems

In formal terminology, the dimension, the coordinate sequence, and the coordinate types are defined by the nature of the particular coordinate system employed. The name, unit scale, unit system, and direction of axis are defined for each axis by the coordinate description entity.

The coordinate sequence is defined according to the axis of the defined datum. If the coordinates are given in a projection, the necessary information is described in the 'entity map projection'. The 'coordinate types' description includes the different height systems which may be used. The range of possible coordinate system types comprises:

Three-dimensional coordinate systems

- Cartesian coordinate system (geocentric);
- geodetic coordinate system with ellipsoidal height (see section 3.1.5);
- geodetic coordinate system with normal height;
- geodetic coordinate system with orthometric height;
- geodetic coordinate system with geopotential number;
- local Cartesian (topocentric);
- astronomical.

Two-dimensional coordinate systems

- geodetic;
- plane;
- local Cartesian;
- astronomical.

One-dimensional systems

- height above ellipsoid;
- orthometric height;
- normal height;
- geopotential number.

A coordinate description always has a name, a definition, and the direction of the axis. An example of a coordinate name is 'E' and of a definition is

'Easting'. A coordinate is necessarily measured according to a system of defined units (e.g. SI units).

3.1.4 Geodetic datum

A tri- or 3-dimensional geodetic datum defines the position of the origin in relation to the Earth mass centre and the orientation of the *Z*-axis in relation to the conventional Earth rotation axis, the *X*-axis to the conventional zero meridian, with the *Y*-axis added in a right-hand system. The parameters of the geodetic ellipsoid are included in the datum definition.

A bi- or 2-dimensional geodetic datum defines the position of a bi-dimensional coordinate system to the Earth body. It is defined by the geodetic coordinates of a main point (the origin of the datum), the deflection of the vertical, and the geoid height at the origin as well as the parameters of the geodetic ellipsoid.

A uni- or 1-dimensional height datum defines the relation of the reference surface of the height system in relation to the mean sea level over the whole ocean (global vertical), the mean sea level as defined by one or more tide gauges (a regional vertical), or to any other preferred reference point.

As indicated earlier, a geodetic datum typically is given a unique name which is a short acronym or meaningful abbreviation of the full name (e.g. ETRS 89 for the European Terrestrial Reference System 1989). The name of the geodetic datum might be the same as the name of the geodetic reference system. The attribute 'full name' should always contain the complete name of the geodetic datum, while the optional description attribute contains a more detailed textual description of the geodetic datum, its recommended usage and possible pitfalls.

To transform coordinates of an older reference system to one of the newer 3-dimensional x, y, z systems is not always a simple task if we do not have the corresponding values for the geoid model. Such uncertainty is surprisingly common: for instance, even in a country as geodetically sophisticated as the United Kingdom, much research was necessary to recover the precise definitions of the 150 or more local, county-related reference systems used in topographic mapping until the 1940s in order to relate historical information to current data. The situation in some other countries is known to be worse with no information being available in certain cases. Two-dimensional ellipsoidal approaches may be easier in such cases though mathematically they tend to be more complex.

3.1.5 The geoid and heights

The Earth gravity field is expressed through its gravity potential which includes both gravitational and centrifugal effects. The gravity g and its spherical directions Φ (astronomical latitude) and Λ (astronomical longitude) are the astronomical coordinates with regard to the CTS.

As defined earlier, the geoid is an equipotential surface of the Earth gravity field which most closely approximates mean sea level. The geoid height N is the positive distance between the ellipsoid and the geoid, defined outside the ellipsoid. In practice, the expression 'closely approximates mean sea level' allows one to consider slightly different equipotential surfaces as geoid models.

Because of multiple geophysical phenomena, the geoid is not a simple mathematical shape although it can be expressed as a sphere with a series of harmonic terms, each one smaller than previously and each one altering the shape to – for instance – a flattened sphere or a pear shape. The larger the number of terms in the expression, the more detail it describes. 'Height' is a number expressing the separation between a point P and a horizontal reference surface. We have already seen the definition of the ellipsoidal height h where the reference surface is a geodetic ellipsoid. But the many types of height normally defined over land typically use the geoid as the reference surface rather than some mathematically defined geodetic ellipsoid. More precisely, they are defined using the geopotential number which is the difference in gravity potential between the geoid and the equipotential surface through the point P and a value for gravity. Geopotential numbers can be converted to linear units. Several options are available and the most commonly used of these are:

- orthometric height H_0, uses a mean value for gravity;
- normal height H_N, uses a more refined gravity value;
- dynamic height H_0, uses an arbitrarily agreed gravity value.

Of these, orthometric heights are most commonly used because they can be determined directly from 'levelling' without first computing geopotential numbers. Orthometric height is simply the linear distance to the geoid, as shown in Figure 3.

The result is that a vertical datum is fully defined by:

- the equipotential surface where the heights are zero or derived by a geoid model. For example, the geoid model used in France for the national geoid

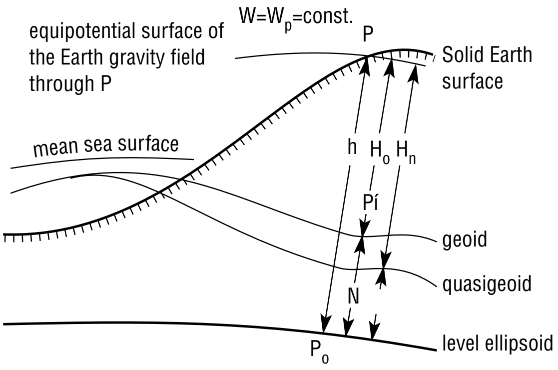


Fig 3. Normal heights and orthometric heights.

system is the equipotential surface passing through the mean level of the Mediterranean sea at Marseilles, determined by tide gauge measurements;

- the type of height (orthometric, normal, dynamic);
- the linear unit in which it is expressed (metre, foot, ...).

The reference system for gravity is given as enumerated values of the gravity system, with the permissible values of IGSN71, Potsdam system and ECS62.

3.1.6 Map projections

Many textbooks provide descriptions of the fundamentals of map projections (e.g. Maling 1973; Snyder 1987, 1993). Put simply, any map projection is a mathematical representation of a geodetic ellipsoid (in some cases, a sphere) as a plane.

Mathematically, if each point *P* on the geodetic ellipsoid is identified by its geographical coordinates (φ,λ) and each point *p*' in the plane by its Cartesian coordinates (E, N) then a map projection is defined by two functions *f* and *g* such that:

$$E = f(\varphi,\lambda)$$

$$N = g(\varphi,\lambda)$$

From this, one can deduce that, for any point *P* in space and having chosen a geodetic reference system, a geodetic ellipsoid, a zero meridian, and a map projection, the position of *P* can be represented by the 3-dimensional coordinates E, N, and h, as shown in Figure 4.

Because of the complexity involved in giving a complete description of all parameters necessary to describe map projections, only well-known projections are recommended here such as Universal Transverse Mercator (UTM) and Gauß-Krüger. For all other projections, the coordinates must be transferred to geographical coordinates if a full description is needed.

3.1.7 Units

As coordinates are given by numerical values (X, Y, Z), (φ,λ,h), (φ,λ), (E, N, h), or (E, N), it is essential to specify the linear and angular units used. Cartesian coordinates (X, Y, Z) are always expressed in metres; heights h and map coordinates (E, N) are in linear units usually given in metres but sometimes

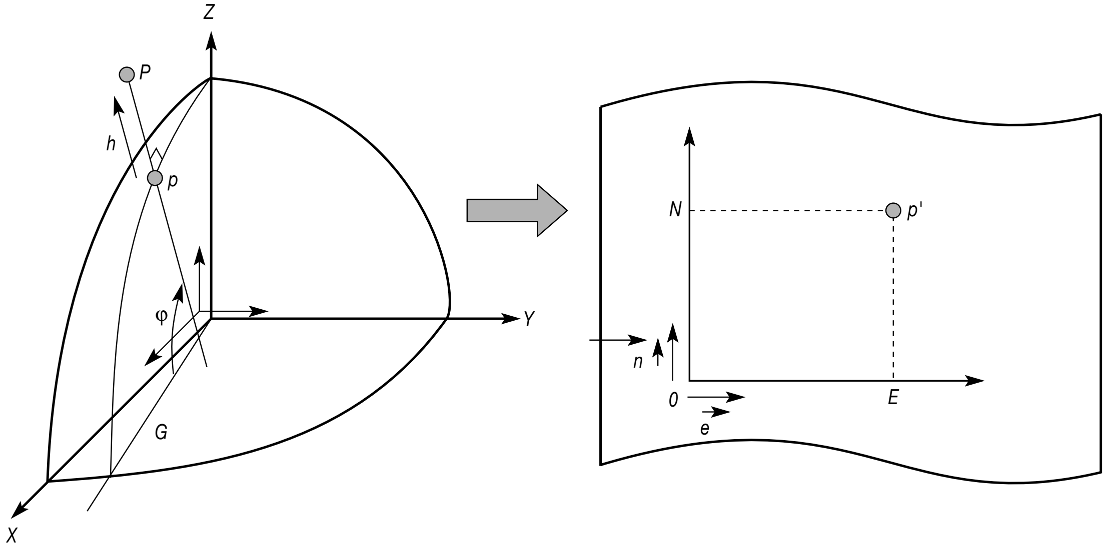


Fig 4. Map projection.

in feet, yards, etc. The geographical coordinates (φ, λ) are usually expressed in angular units (e.g. sexagesimal degrees, minutes of arc, and seconds of arc, but also in gons (grades or centesimal degrees), radians, or decimal degrees.

Local Cartesian coordinates (u, v, w) are defined as Cartesian coordinates of a local reference frame, such as those shown in Figure 5 which define an affine frame around a point P at the Earth's surface. Its orientation with regard to the CTS defines it completely.

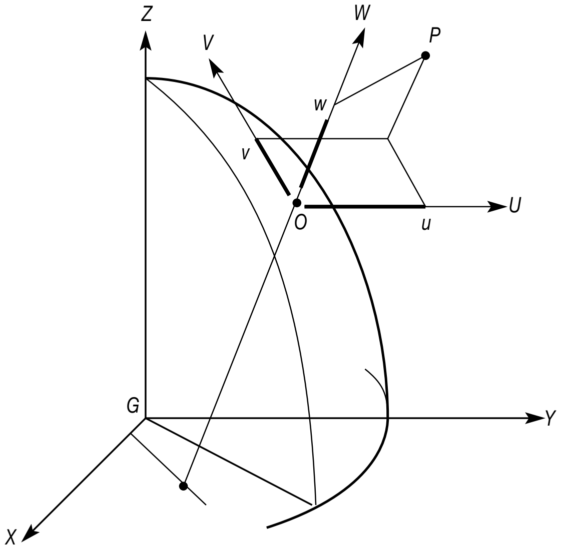
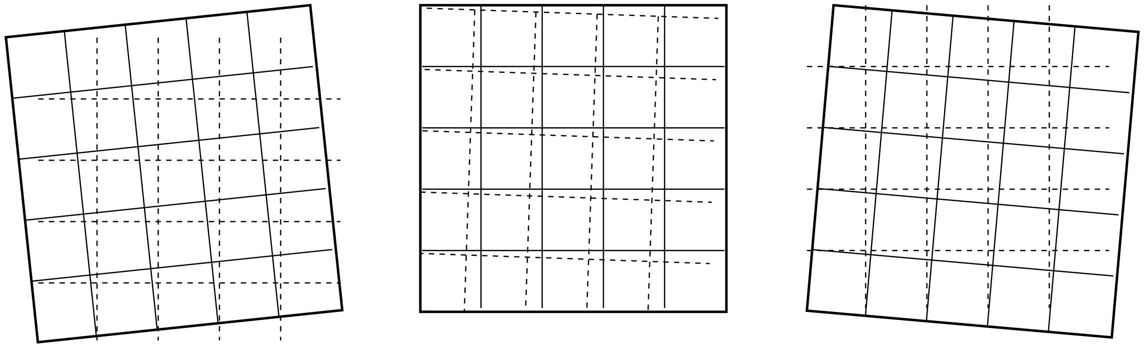


Fig 5. Local Cartesian coordinates.

3.1.8 National versus global reference systems

The principles and situations described above can be summarised as follows. For reasons of history and the available technology, conventional geodetic coordinates, which are expressed in terms of geodetic latitude and longitude, are related to a national or a continental datum. Thus, for example, the UK Primary Triangulation coordinates are referred to as the OS(SN)80 Datum which is defined by the Airy ellipsoid and the adopted coordinates of the Origin pillar in the grounds of Herstmonceaux Castle in Sussex in southern Britain. Similarly, coordinates in the North Sea prior to the widespread use of GPS were expressed in the ED50 Datum and are related to the Hayford ellipsoid and a point of origin in Germany. Finally, different map projections or different parameters of the same map projections are in use in many countries, each tailored to maximise or minimise some properties held to be of high importance (e.g. distortion; see Maling 1973, Chapter 4). The situation is exacerbated where discontinuities occur in the map projections used, notably in the widely used UTM projection. Britain, for instance, straddles UTM zones and there is therefore an obviously variable relationship between the National Grid coordinates familiar to most of the populace and those of UTM – as illustrated in Figure 6. Where all types of data are held in computer form within one projection system, transformation to another may cause other visual problems; this is most acute where the data are held in raster form, as illustrated in Figure 7 (see also Dowman, Chapter 31).



NG4050SW (W Scotland, zone 29) NJ5050SW (E Scotland, zone 30) TG5000SW (Norfolk, zone 31)

Fig 6. Ordnance Survey 1:10 000 scale map tiles transformed into WGS coordinates and projected onto the UTM zones in which they fall.

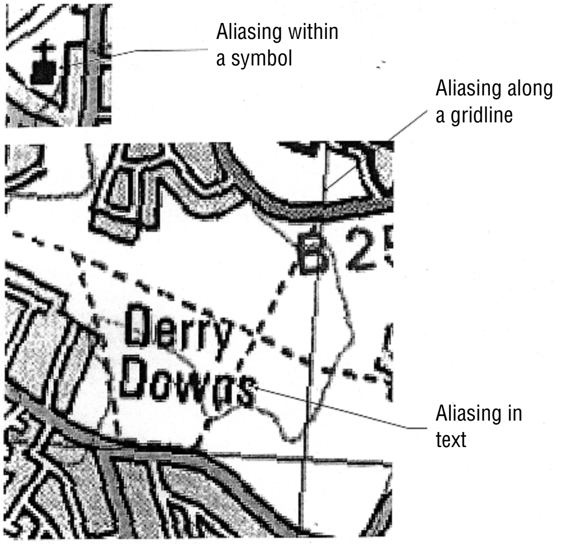


Fig 7. Raster aliasing effect attributable to transforming data encoded in one reference system and map projection combination into WGS coordinates and reprojecting these onto UTM. The effect could be mitigated using smoothing and interpolation techniques.

3.2 Indirect positioning

Common usage of geographical description includes narrative where position is described by place name and relationships to that place (e.g. '25 kilometres north west of Frankfurt'). Alternative indirect ways of describing geography include post (or zip) codes (see Martin, Chapter 6), a variety of zoning systems (such as the French départements), and distances through a network. Some geographical analysis has even been carried out using only the knowledge of the adjacencies between different zones. Although some of these means of describing position are commonplace, they cannot be used to link together different datasets (unless the different datasets each pertain to the same named entities). Spatial relationships cannot be computed nor can the data generally be mapped in geographical space.

It follows that to carry out almost any GIS operation requires that an added value process be carried out: these locations must therefore be described by coordinates of some type, as described in section 3.1. Practical difficulties often occur in this process for the bounds of some entities used in common parlance (e.g. 'the Alps' or 'the Corn Belt') may be fuzzy to some degree (see Mark, Chapter 7, for a general discussion of this topic). Nevertheless, an increasing number of organisations now maintain and offer mechanisms and data to enable such added value to be achieved. The most common translations of this kind are to relate post (or zip) codes to coordinates or even individual postal addresses within a post (or zip) code to coordinates (as in the British ADDRESS-POINT product).

4 CONVERSION OF DATA REFERENCED TO DIFFERENT POSITIONING SYSTEMS

In recent years positioning based on GPS-observations has proved more and more powerful, almost totally replacing classical triangulation techniques. Thus modern reference systems such as the WGS84 or the ETRS89 are used increasingly in the fields of practical surveying. In so doing, one must be very careful not to generate errors by mixing reference systems and datums. The easiest way to anticipate any difficulties is usually to contact the national mapping organisation for advice in the first instance. Alternatively, it is possible to make calculations of the interrelationship through use of a certain number of stations identical in both the national and the satellite systems (ideally five or more). As positioning through use of navigation satellite systems (GPS, GLONASS, PRARE etc.) will grow rapidly in the future, special attention has to be paid to the different aspects of using various reference systems in daily surveying and, even more important, in the use of data compiled from different sources inside GIS.

The primary reason for this is the historical legacy. Most geographical data in current use have been compiled on the basis of national reference systems which predate the new global system. These reference systems are embedded in national map series at all scales. While some countries (such as Australia and the USA) are committed to converting their maps series to being based on WGS84 or ETRFS, even in these cases the transition period is likely to be long and the cost measured in many millions or billions of dollars. In addition, many users have already expended considerable investments in encoding their own data in relation to these national systems.

This would be less of a problem if everything could simply be converted purely algebraically without loss of precision or accuracy and all those holding data in relation to national frameworks used the same national conversion parameters. In reality, the problem is much greater than this (see Salgé, Chapter 50). In the first instance, there is a huge number of different combinations of geodetic reference system, geodetic ellipsoid, zero meridian, and map projection in use and (surprisingly perhaps) records of which ones have been used are sometimes not kept. The basic work on the geodetic system may, for instance, have been carried out 50 or more years ago, before computers were available to support adjustments in the geodetic networks. Typically, such networks were assembled on a piece-by-piece basis although overall checks were made. As a consequence of this, of generalisations and errors built into maps based on these networks, of the loss of recorded attribute information when digitising the maps and the widespread, free availability of variable quality approximations to the overall algebraic transformations, conversion of existing data into the WGS framework has proved much more troublesome than is generally appreciated.

Ashkenazi (1986) anticipated many of these problems and has demonstrated their significance in subsequent papers. He has shown that – even without errors in the maps or in their digitising – coordinate shifts of the order of several hundred metres can occur between the position of locations as described in WGS84 and pre-existing national frameworks. Even the simple act of transforming from one map projection system to another may cause considerable difficulties, especially in raster topographic datasets (see Figures 6 and 7).

5 CONCLUSIONS

The routine use of GPS receivers or GIS as 'black boxes' can foster a view that 'truth' is being constructed or handled, and that it may be compared directly with other respected and geographically precise sources. This chapter has set out to demonstrate that, even in a relatively well-defined (but crucial) element of GIS, many complexities and gross errors may occur unless some fundamental principles are understood by the user of the system. Thus care in the selection and use of referencing and coordinate systems is always essential, especially when the need to link different datasets together requires use of information defined and collated in different systems.

The most common difficulties arise when information derived by the use of GPS is combined with that from maps produced using different datums. The combined effects of the use of different geodetic reference systems (including different datums), projection change, grid shifts, and any embedded errors in the maps can cause great practical problems. As indicated earlier, maps are the basis for much current and most historical data in GIS so this problem has to be faced. It can only be faced with a good understanding of the science of geodesy and a good knowledge of the genealogy of the mapping from which data have been derived.

References

Ashkenazi V 1986 Coordinate systems: how to get your position very precise and completely wrong. *Journal of Navigation* 39: 269–78

Bomford G 1980 *Geodesy*, 4th edition. Oxford, Clarendon Press

Hofmann-Wellenhof B, Lichtenegger H, Collins J 1994 *Global positioning systems: theory and practice*, 3rd edition. Berlin, Springer

Kennedy M 1996 *The global positioning system and GIS: an introduction*. Michigan, Ann Arbor Press

Leick A 1990 *GPS satellite surveying*. New York, John Wiley & Sons Inc.

Maling D H 1973 *Coordinate systems and map projections*. London, George Philip

Snyder J P 1987 *Map projections: a working manual.* US Geological Survey Professional Paper 1395. Washington DC, USGS

Snyder J P 1993 *Flattening the Earth: two thousand years of map projections*. Chicago and London, Chicago University Press

Vanicek D, Krakovsky E D 1986 *Geodesy: the concepts*. Amsterdam, New Holland

Warita Y, Nonomura K 1997 The national and global activities of the Japanese national mapping organisation. In Rhind D (ed.) *Framework for the world.* Cambridge (UK), GeoInformation International: 31–47