31

Encoding and validating data from maps and images

I J DOWMAN

This chapter gives an overview of how data for a GIS can be derived from existing graphical information (a map) or from image data such as aerial photographs and satellite images. Two crucial principles underlie this process: (a) the need to understand the characteristics of the original data, including the quality of the source data; the processes used to produce the document to be encoded; and the geometric characteristics of the data; (b) the need to understand the processes used to produce the encoded data and their subsequent quality.

In order to obtain this understanding, this chapter gives some detail of the characteristics of the data sources and of the methods of converting the original data into a form suitable for a GIS. The need to geometrically correct image data leads to the use of elevation data to correct for relief effects and therefore the principles of using stereoscopic data are outlined and methods for deriving digital elevation models. It is assumed that automation will be used in these processes as much as possible, so an outline is given of the automated tools which are now in use. Finally, the question of accuracy is discussed, including the factors which affect accuracy, how accuracy can be assessed, and how this information can be given to and applied by the user of a GIS.

1 INTRODUCTION

Many of the data which are used in GIS come from existing graphical products printed onto paper, or from images derived from photographic cameras or digital sensors (e.g. see Bibby and Shepherd, Chapter 68, Waters, Chapter 59). Some of the latter comes from sensors carried on satellites. In order to use these data in a GIS they must be processed to be in a form which is compatible with the GIS. The first point to consider is whether the GIS requires raster or vector data – assuming that the GIS cannot convert between the two without imposing undesirable constraints. The second point concerns the reference system: all data which are to be overlaid, merged, or compared must be in the same map projection and on the same datum. This implies that any distortions present because of the data capture system must be corrected. The third point is that the quality of the data must be known (see Veregin, Chapter 12). This chapter is concerned with how vector or raster data can be acquired from existing maps or from images and how they must be processed in order to be used in a GIS. In order to carry out these processes it is essential that the characteristics of the original data and the processes used in the preparation are understood. It is also essential that all of the possible sources of error are recognised, particularly with regard to how such errors affect the quality of the final dataset (see also Fisher, Chapter 13; Heuvelink, Chapter 14).

The term 'map' is used in this chapter to refer to a dataset which contains accurate information about features on the surface of the Earth. In the past there would have been no ambiguity about this because a map user would be handling a piece of paper with the required information printed onto it. Now this information may also be stored in digital form (see Longley et al, Chapter 72, for a discussion of the 'map metaphor').

A knowledge of map projections is essential when handling data from different sources. Any map

projection will introduce some distortion into the transformation from the ground to paper and it is important that the sources of this are understood. Data mapped using different projections cannot generally be combined and processes of conversion and transformation must therefore be applied. In addition different datums may be used, even on the same projection, and these too must be understood. A detailed discussion is beyond the scope of this chapter but reference may be made to Snyder (1986; see also Barnsley, Chapter 32; Seeger, Chapter 30).

The first step in using map data in a GIS is to encode it into vector or raster format. This involves either line scanning or raster scanning. Because of the monotony of this task, it is desirable that as much automation as possible is introduced. After encoding, the data must be transformed into the required projection and any known errors removed. A raw image derived from a sensor in an aircraft or on a satellite will contain distortions attributable to the attitude of the sensor and the relief of the ground, and it will not show all the features which a user will need (Barnsley, Chapter 32; Meyers, Chapter 57). Such distortions must be corrected before data derived from the images can be used in a defined reference system. Correction may be possible through 2-dimensional transformation, although 3-dimensional correction using stereoscopic imagery or digital elevation models (DEMs) is frequently also necessary. Such processes are part of the discipline of photogrammetry and this section can be regarded as an introduction to that subject.

A vast amount of useful data already exist in the form of printed maps and an immediate response to the requirement for digital data is to assume that these existing data should form the basis of the digital database. It must, however, be remembered that a great deal of effort, and hence cost, is required to convert printed maps to digital data. A figure of 60–80 per cent is often given, without substantiation, as being the proportion of the cost of establishing a GIS which is attributable to obtaining the data. It must also be remembered that maps may be out of date and may have a number of errors inherent in them, and that the conversion process will itself add further errors. Before embarking on a data acquisition exercise it is sensible to weigh up all the factors involved in data collection, and to consider the available sources of data in the context of the use to which the data will be put and the methods which will be used to update the data. Logan (1995) provides a useful discussion on these topics.

2 ENCODING MAP DATA

2.1 Principles

Map data have two principle components: geometry and attributes. For use in a GIS every point in a dataset must have unique coordinates (x and y or x, y, and z) and each must be labelled. On a printed map the coordinates usually come from a grid or graticule on the map, provided in such a way that the (x, y) coordinates of any point can be read off a scale. The information about what the point represents will come from the context or the way in which the point is represented, and this representation can be converted to an attribute through a legend. When converting from a printed map to a digital database, both of these components must be recorded.

Digitising will always take place in a rectangular coordinate system which is defined by the digitising system. Conversion from this to a national or regional grid system is therefore relatively straightforward. The procedure is usually to record the corner points of the document to be recorded in the digitising system and then to carry out a transformation (see section 5.3) between the recorded coordinates and the known coordinates in the national system of the corner points. An affine transformation in often used for this to correct for any errors created by paper distortion. Digitising is often carried out in patches (or tiles) which may be the size of a grid square or larger and each tile is transformed to the coordinates of its corners. It is always wise to carry out checks with well-defined points of detail on the map.

Attributes must be assigned by the operator, or automatically where possible, through the use of an attribute menu or key coding.

2.2 Equipment

The most straightforward form of conversion is to digitise manually the information from the printed map into either point, line, or area features and to assign each feature an attribute code. The features are then recorded as coordinates in point or vector form. This can be done on the digitising table on

which the map is placed. The operator then selects points or trace lines and the coordinates are recorded as the cursor moves over the map. After digitising, the result needs to be carefully checked and edited to ensure that everything has been recorded and that lines join up where they should and do not overrun. If carried out systematically and carefully this method can result in accurate and complete vector data. The disadvantage is the time which is required and the tedious nature of the task.

A map can be quickly and easily recorded automatically in digital form using scanning techniques. However these methods only record the data in terms of position in a raster format and do not link to separate attribute files, since there is no feature coding or topology. Scanning can be accurately carried out on drum scanners, whereby the printed document is fixed to an accurately rotating drum. The drum is scanned by an optical reading head which determines whether a line or symbol exists over a 'pixel' (picture element) of a given size – which can be as small as 6μm. Colour documents can be recorded by scanning three times, in the red, green, and blue regions of the spectrum. The result is a raster image which reconstructs the original as pixels rather than as lines and symbols and which may look the same as the original but cannot be interpreted by a computer. The accuracy and visual impression is controlled by the pixel size, but a smaller pixel size means more data, more time to scan and display, and increased storage requirements. Current problems with scanning are centred on accurate colour reproduction and the size of files which will be produced for large documents. The former problem can be overcome at the expense of greater data handling if colour separates are available.

Less expensive but less accurate desktop scanners, designed for the publishing trade, are available which will scan at up to 1200 dots per inch (giving a pixel size of 20.16μm). Desktop scanners are subject to geometric distortion but this can mostly be removed by calibrating the scanner and subsequent correction (Sarjakoski 1992).

2.3 Raster-to-vector conversion

The raster image can be converted to vector by a raster-to-vector conversion. This technique is not fully developed as an automatic process and still cannot add attributes. It also requires extensive editing to remove text and symbols. Raster-to-vector conversion requires salient pixels to be assigned (x, y) coordinates. The actual coordination of a pixel is no problem but deciding which pixel should represent a corner or a junction of two lines is not easy. A junction, for example, may be represented by several pixels and the lines forming the intersection may be several pixels wide; the intersection itself must, however, be represented by a single point. Line thinning will reduce line size but an algorithm is needed to select the optimum point to represent the intersection. This is illustrated in Figure 1, where the node of the intersection could fall at any of the four pixels within the small box. Features such as roads may be represented as single lines or as parallel lines, and the decision as to which to use and how to determine which lines to digitise can also be difficult for an automatic system.

The most efficient solution for raster-to-vector conversion at present may be semi-automatic systems. These are based on line-following techniques which can be done quite well by a computer. The operator is better at starting the process and making decisions about which way to go at intersections of lines. The VTRAK system of Laser-Scan Limited (Cambridge, UK) is a good example of an automatic line-following system: it

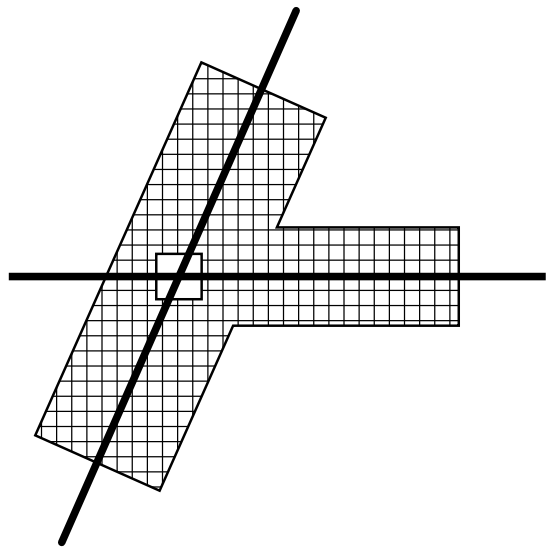


Fig 1. An example of a problem in raster-to-vector conversion.

can be used to digitise features into a range of databases, notably object-oriented databases. It also provides editing and quality control facilities and automatic text recognition.

2.4 Editing and error analysis

Editing is an important process in any digitising operation because errors will always be made. The most common error is poor closing of lines at intersections or in polygons. Most digitising packages now include efficient tools for editing data (see Bernhardsen, Chapter 41). Figure 2 shows some examples of common errors.

Errors in content have been discussed above and can be detected by careful checking. Errors in geometry are not so easy to detect. The most important rule in checking accuracy is to have otherwise redundant check points. As explained above a transformation is necessary between the digitising system and the reference system. Usually four points are a minimum requirement to carry out a transformation. A number of additional check points should also be used and the coordinates of transformed points in the digital document should be compared with the original values. Some useful validation statistics are discussed in section 8. A further useful check is possible if other data of the same area exist and the two sets can be superimposed.

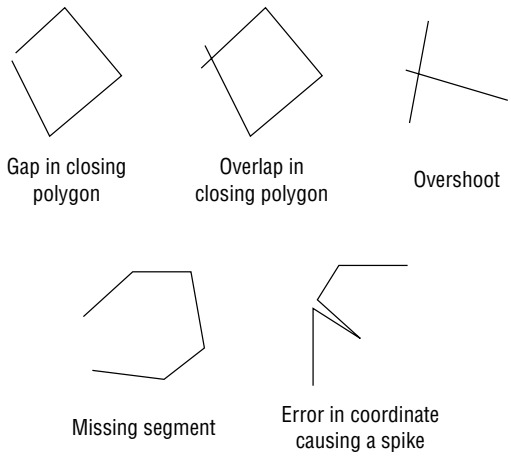


Fig 2. Common errors in digitising.

3 THE GEOMETRY OF AERIAL PHOTOGRAPHS AND SATELLITE DATA

3.1 Image formation

In order to produce a map from any satellite image it is necessary to first define the geometry of the image forming system, to consider the movement of the platform on which the sensor is mounted, and to define the shape of the ground which is covered by the image. The importance of each of these factors will depend upon the type of sensor used and the path which the sensor is following relative to the ground. This section is concerned with the general effects and these are related to different sensor types. Specific sensors for mapping are discussed in section 4 and by Barnsley (Chapter 32) and Estes and Loveland (Chapter 48).

Images from a particular sensor are usually presented in a standard format covering a defined area of ground. Such an image is called a frame. A frame may be formed in three ways:

1. As a single exposure – that is with no significant movement of the sensor whilst the image is formed as is found in the case of a frame camera. This is used for central projection which is used in all aerial photography and in a number of satellite sensors.
2. As a series of lines almost normal to the track of the sensor. In this case time must be considered in the model for defining the construction of a full frame. The main types of sensor in this category are the push broom scanners, of which SPOT is the best contemporary example.
3. As a series of points each recorded at a separate time. This is the most distorted type of image requiring the most complex mathematical model. The scanner systems such as LANDSAT fall into this category.

The photographic camera, in which an image is formed by light rays from an object being focused onto a focal plane by a lens, may be taken as the standard case. This lens acts as a point through which all light rays pass, so that the process of image formation can be seen as a series of straight lines passing through the perspective centre S onto the focal plane. If the objects all fall on a plane which is parallel to the focal plane, then a direct reduction of the object is found in the focal plane and the amount of reduction is given by the ratio of the object

distance to the image distance. This is an approximation to the case of vertical aerial photography when the ground is the object, the object distance is the flying height *H,* and the image distance is the principal distance of the camera *f* (equal to the focal length of the camera when the object distance is effectively infinity). Two points are defined on the vertical line passing through *S* as the nadir points on the image (*n*) and on the ground (*N*). The scale of the photograph in this case is *f*/*H*. This is illustrated in Figure 3. In practice the ground is not flat and the focal plane is not horizontal.

3.2 Effect of relief

The effect of relief is to cause the scale to change throughout the image and to cause images to be displaced, radially from the nadir, from the position at which they should appear if there were no relief. In Figure 4 a point *A* on the surface of the ground should appear on a map of scale *f*/*H* at *a*' but actually appears on the photograph at *a*.

The magnitude of this displacement is given by the expression

$$aa' = an \cdot \frac{h}{H} \qquad (1)$$

or in terms of a radial distance from the nadir points (*r*)

$$dr = r \cdot \frac{h}{H} \qquad (2)$$

This expression indicates that the effect of relief displacement is less from greater sensor altitudes. For example on an aerial photograph taken from 10 000 metres, relief of 500 metres will cause a displacement of 5 mm for a point appearing on the edge of a photograph (r = 100 mm). This is equivalent to 660 metres on the ground if a wide angle camera is used. On an image taken from a satellite pointing vertically downwards, with an altitude of 700 km the effect of relief on the ground will be 64 metres. For normal relief the effect on this type of satellite image can be ignored, but for photographic cameras and tilted satellite sensors it cannot be ignored and 3-dimensional geometry must be considered. In Plate 19 the effect of relief distortion can be clearly seen at the edge of the photograph where buildings appear to lean outwards. In other words the top of the building is in a different position to the bottom.

The effect of relief can be removed by using a rigorous 3-dimensional model of the geometry of two images or by having a digital elevation model available in order to compute, and hence correct, the effect of relief at every point. An image which has been corrected for the effect of relief is called an *orthoimage*.

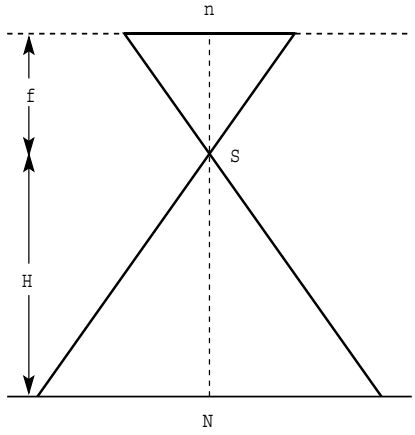


Fig 3. The scale of an image.

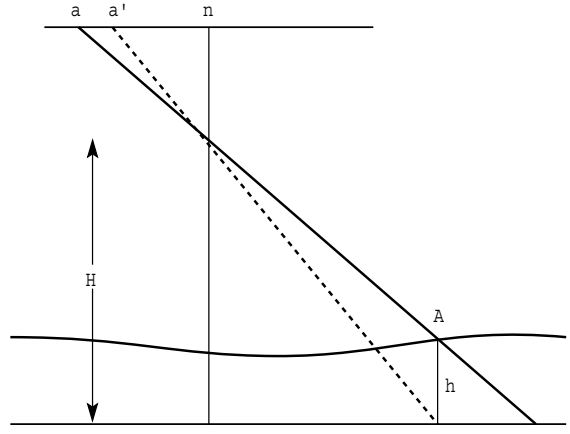


Fig 4. Distortion attributable to relief.

3.3 Distortion because of tilt and other causes

There are other distortions in imagery which must also be considered. These include the tilt of the sensor platform, the movement of the platform and the movement of the Earth. Most of these are small, although they still need to be removed if accurate data are required. A source of distortion which is potentially very serious arises if the sensor is tilted through a large angle – for example SPOT images can be tilted by up to 27 degrees and the new high resolution sensors will tilt by up to 45 degrees. This is an effect which must be removed for accurate mapping. The methods of correcting distortion are described in section 5.

3.4 Acquisition of stereoscopic data

If images from two different positions are available, then the pair of images can be viewed stereoscopically to give a 3-dimensional view of the terrain and measurements can be made to give 3-dimensional coordinates. In order to obtain a stereoscopic image from a pair of photographs certain conditions must be satisfied. The cameras must neither be too close together, nor too far apart. If they are too close together the view will not be different enough to enable a stereoscopic image to be formed; if they are too far apart the views will not be similar enough. Photographs should also be taken with the same or similar cameras and be taken from approximately the same distance from the object. The basic geometric condition can be expressed in terms of base to height ratio.

In Figure 5 the base is shown to be the distance between the cameras, B, and the distance between the base and the object is the height. In aerial photography the distance is equivalent to the flying height, H, and we have a base to height ratio, $B{:}H$. In order to obtain good stereoscopic viewing $B{:}H$ should lie between the limits 0.3 and 1.0.

Photographs taken with a camera from an aircraft for the purpose of constructing a map are carefully controlled so that the axis of the camera is pointing almost vertically downwards and each photograph overlaps its neighbour by 60 per cent. This ensures complete coverage of the ground and a convenient base to height ratio. This is shown in Figure 6. Stereoscopic coverage from satellite sensors is more variable and because of the great altitude of the sensor coverage from vertical pointing sensors does not usually give a suitable base to height ratio for mapping.

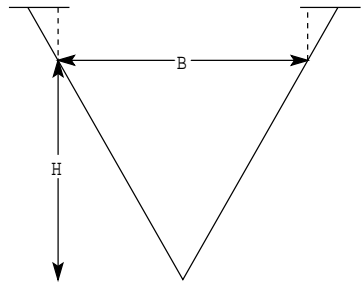


Fig 5. Base to height ratio.

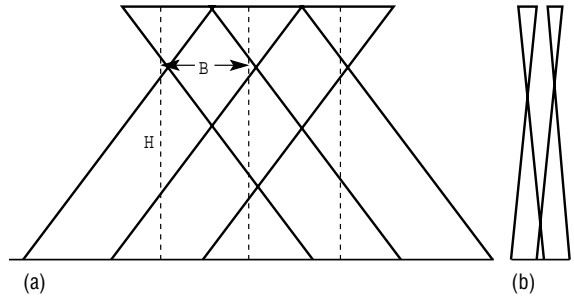


Fig 6. Base to height ratios of: (a) aerial photography; (b) vertical satellite images.

3.5 Heights from stereoscopic data

As already noted, stereoscopic images can be used for measuring heights. The principle is shown in Figure 7. On two images, two points of different heights, A and B, will produce images at a_1, b_1 on the left hand photograph and at a_2, b_2 on the right photograph. The separation of the images $a_1p_1 + a_2p_2$ and $b_1p_1 + b_2p_2$ is clearly proportional to the heights of A and B. These separations are called parallaxes or disparities.

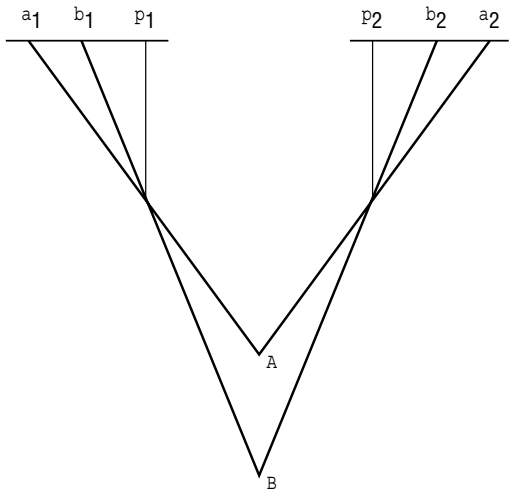


Fig 7. The principle of height determination.

Using the principle as shown in Figure 7 we can derive an expression which relates the height of a point *(Z)* to the camera geometry and the parallax *(p)*

$$Z = \frac{f B}{p} \tag{3}$$

and considering a small change in height (*dZ*) we can show that:

$$dZ = \frac{Z}{f} \cdot \frac{Z}{B} dp \tag{4}$$

Equation 4 is an expression which shows how a small change in parallax (*dp*) will give a small change in height (*dZ*). This equation can be used to give the precision with which height differences can be found. Precision is usually measured by the standard deviation of a set of measurements, or quantities derived from measurement. In this case the standard deviation of height determination, s_z can be found if the precision of observation of parallax s_p is known and the equation becomes:

$$s_z = \frac{Z}{f} \cdot \frac{Z}{B} s_p \tag{5}$$

which gives an expression showing that the precision with which height can be measured is related to the scale of the photograph (*Z*/*f*), the base-to-height ratio (*Z*/*B*), and the precision of measurement of parallax.

4 IMAGE ACQUISITION SYSTEMS

Images for mapping may be obtained from sensors placed in aircraft or on satellite platforms. Airborne sensors are mainly film cameras producing high quality images suitable for photogrammetric mapping. Sensors from satellites generally produce images directly in digital form as the satellite moves forwards: the movement enables the necessary large volumes of data to be recorded, but complicates the image geometry. A summary of the main sensors is given in Table 1. The characteristics of satellite sensors are discussed in more detail by Barnsley (Chapter 32) and Estes and Loveland (Chapter 48). Here only a few of the sensors which can be used for mapping will be discussed. Film cameras are used in space and the Russians are the main operators of such cameras. The KFA1000 and KVR1000 are two examples which are included in Table 1. Both of these are designed to give large scale images rather than height information.

Scanners are the most widely used imaging systems from space, and LANDSAT is probably the best known. A scanner records a single pixel (or a group of pixels) at a time; a strip approximately at right angles to the direction of flight is imaged as a mirror rotates about the axis of flight. The raw data obtained from a scanner are subject to many distortions which must be corrected prior to use within a GIS, but because many of the sensors point only directly downwards, rigorous correction is not required.

Pushbroom sensors are the most important for mapping. SPOT was the first of these and was designed as a mapping sensor: it is still dominant in this area, although other systems have been designed and flown. The US high resolution sensors are an important new form of pushbroom scanner.

Stereoscopic overlap can be arranged either fore and aft or laterally so that 3-dimensional models can be formed. Displacement attributable to the relief of the ground is normal to the direction of flight, that is along the arrays, and the amount of displacement is proportional to the distance from the principal point and the altitude. If lateral overlap is used by tilting the sensor than the effect of relief will be large, but a good base to height ratio will be obtained for height measurement.

The SPOT satellite uses a 'lateral overlap' sensor. This gives a good base-to-height ratio but suffers from the disadvantage that stereoscopic images will

Table 1 Characteristics of sensors used for mapping (see also Barnsley, Chapter 32 Table 2).

Platform	*Sensor*	*Launch date*	*Type*	*Pixel size (m)*	*Swath width (km)*	*B:H*	*Height accuracy (m)*
Aircraft	Film camera		Principal manufacturers are Zeiss and Leica		†	0.3 – 1.0	Depends of flying height: 0.01% overhead
Cosmos	KFA1000	1994	Russian film camera, f = 1000mm	5–10*	†	0.12	30
Cosmos	KVR1000	1987	Russian panoramic film camera, f=1000mm	2*	160	0	N/A
Landsat	Thematic Mapper	1982	Scanning system	29	180	N/A	N/A
SPOT 1-4	HRV	1986–97	Cross track push broom	10,20	60	≤1	10
JERS-1	OPS	1992	2 sensor along track stereo	20	75	0.3	50
Priroda	MOMS	1996	3 sensor along track stereo	6	40	0.9	3–6
IRS-1C	Pan	1995	Cross track stereo	5.8	70	1	5–10§
Earlybird		1997	Along track stereo	3	6	1.2	1.5‡
Quickbird		1997	Along track stereo	1	36	1.2	1.5‡
Space Imaging		1997	Along track stereo	1	11	2	0.4‡
Orbview		1998	Along track and cross track stereo	1–2	8	2	0.4‡

† Varies with altitude
§ Preliminary result
* Originally photographic film but available in digital format.
‡ Predicted result

be taken a minimum of one day apart – usually longer, as cloud conditions may change from one day to the next. 'Along track' stereo imaging is more efficient as it is very likely to generate stereoscopic data. The optical sensors (OPS) on JERS-1 and MOMS-02 are along track systems: other systems have been designed and will be implemented in the future. Of particular note are the high resolution sensors such as Earlybird and a system from Space Imaging, being developed by commercial organisations in the USA. An example of simulated data is shown in Plate 18, which is part of Space Imaging's CARTERRA San Francisco series of high resolution, high accuracy imagery. It was derived by fusing a one metre resolution panchromatic image and a four metre resolution multispectral image. Laid on top of a United States Geological Survey (USGS) seven-and-a-half minute quad sheet, this image illustrates the value of high resolution imagery in place of traditional line-drawn maps. They are used for a variety of applications, including infrastructure management, urban planning, utilities, and transportation.

Another type of sensor is synthetic aperture radar (SAR). Images from these sensors have a quite different geometry and appearance to those of other sensors and details can be found in specialist literature on the subject (Schreier 1993). Radar is becoming particularly important at the moment because of its potential to produce high accuracy elevation data through SAR interferometry.

5 CORRECTION OF ERRORS IN IMAGES FOR USE IN GIS

5.1 Requirements

The quality of image data to be used in a GIS must match that of the use of the data. Although many applications require only 2-dimensional data, 3-dimensional information may nevertheless be needed to apply the necessary corrections. There are two main methods of correction. First, a 2-dimensional transformation which is suitable for application to images in which there is very low relief distortion – as, for example, in LANDSAT data. Second, a rigorous 3-dimensional correction which corrects for all the effects of relief, sensor orientation, Earth rotation, and Earth curvature.

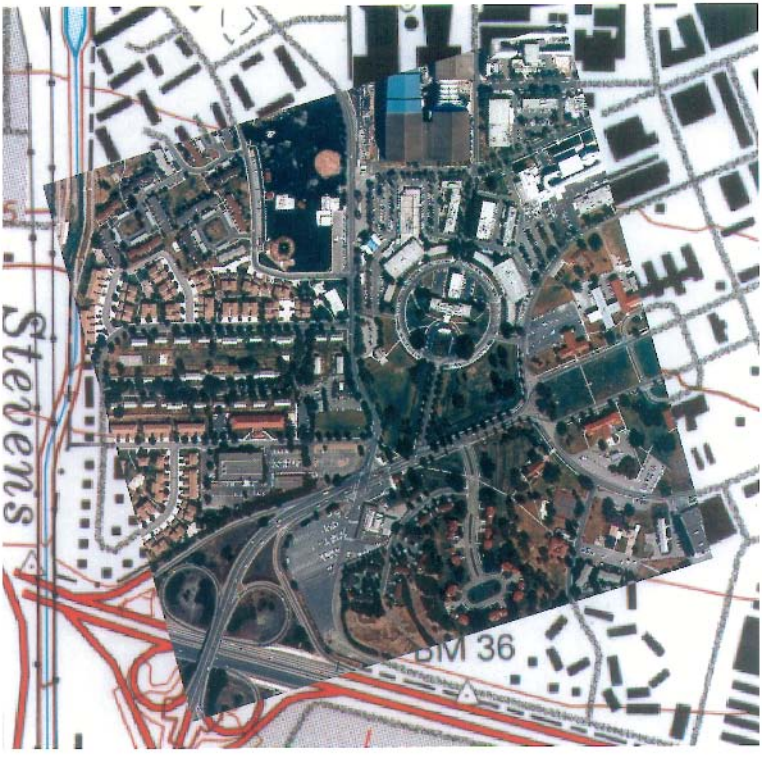


Plate 18
High resolution satellite data of Mountain View, California.
(*Source*: Space Imaging 1996)

For both of these methods ground control points (GCPs) are generally required.

5.2 Ground control

A GCP must be recognisable in both images (or one image if stereo is not being used) and have known coordinates in a suitable ground reference system. The number of points required depends on the method used (Seeger, Chapter 30). GCP coordinates may be obtained directly by survey measurement or from a map. In either case the coordinates will be given in a reference system; this may be geographical (latitude and longitude) or Cartesian (x, y, z) and it may be global, based on the centre of the Earth, or local, based on a national or regional projection. It is always important that the characteristics of the reference system are known and that all coordinates are given in the same system. Further information of reference systems and conversion between systems may be found in Seeger (Chapter 30).

Direct survey measurements may come from surveys based on a local or national coordinate system (e.g. the UK's National Grid), or they may come from the Global Positioning System (GPS) which allows coordinates to be fixed directly from navigation satellites (Lange and Gilbert, Chapter 33). Maps should be used with caution for determining GCPs. Map data at scales of 1:25 000 and smaller are notoriously unreliable because of the many errors which may have accumulated in the map production and map digitising process. These include survey errors (in some parts of the world published maps may be based on topographical sketches), drafting errors, generalisation, paper distortion, and errors in digitising a paper document for use in the validation process. While it is always necessary to take the accuracy of the data into account when using ground control, it is particularly important when using map data.

5.3 Plane transformations

5.3.1 Introduction

For areas of low relief or for low resolution sensors, fairly simple methods of correction may be used. The correction of data in two dimensions may be approached by applying a transformation to the data and resampling the data to produce a corrected image which gives a best fit to the ground control used. The transformation may be based on a theoretical consideration of the errors involved or may be selected on empirical grounds. The method is the one most commonly used to produce an image which is corrected to fit to a given map projection.

A number of transformations are widely used and a brief description of the common ones is given here. An image coordinate system (x, y) and a ground coordinate system (X, Y) is assumed.

5.3.2 Two-dimensional similarity transformation (four parameters)

This transformation is to relate any 2-dimensional rectangular coordinate system to any other 2-dimensional rectangular coordinate system. It preserves the internal geometry of the transformed system, so it is ideal for comparing the geometry of any two systems simply by determining the residuals and the root mean square errors after transformation. For a given control point, the two equations:

$$X = ax - by + c \tag{6}$$
$$Y = bx + ay + d$$

may define this transformation.

A similarity transformation is performed by applying a scale factor ($m = (a^2 + b^2)$), a rotation angle ($tan\ \alpha = b/a$) and two translations (c and d). A minimum of two GCPs is needed although it is always desirable to have more for purposes of checking.

5.3.3 Two-dimensional affine transformation (six parameters)

A mathematical relationship for an affine transformation may be expressed in the following:

$$X = a_o + a_1 x + a_2 y \tag{7}$$
$$Y = b_o + b_1 x + b_2 y$$

An affine transformation enables adjustment to be applied independently in each direction. Thus for scanner images it corrects first-order distortions such as affinity attributable to non-orthogonality and scale difference between scan along track directions which may be caused by earth rotation, map projection, and other geometric distortions. A minimum of three GCPs is required.

5.3.4 Second-order polynomials (twelve-parameters)

Polynomials in the form:

$$X = a_o + a_1x + a_2y + a_3x^2 + a_4y^2 + a_5xy \quad (8)$$

$$Y = b_o + b_1x + b_2y + b_3x^2 + b_4y^2 + b_5xy$$

are used for correction of scanner data. If polynomials are used, great care must be taken to ensure that a sufficient number of control points is available and that they are distributed over the whole area to be transformed. A minimum of six GCPs is necessarily required to determine the transformation parameters, although it is desirable to have more to build in checks. In addition to first-order distortions, polynomials correct second-order distortions caused by pitch and roll, subsatellite track curvature and scan lines convergence because of Earth rotation and map projection. They may also correct some of the distortions related to the attitude variations along the flight path.

Additional terms may be added to Equation 8 to correct for higher order distortions. The need for care in use of control points is greater for higher orders.

5.3.5 Resampling

After calculating the parameters of the transformations, the transformation must be applied to the image and new Gray level values be computed for each pixel in the operation of resampling. Resampling may introduce some changes to the data and distort the characteristics of some features.

5.4 Determination of orientation elements

The process of extracting 3-dimensional coordinates must include the determination of the orientation and position of the sensor and the application of this information to the measured image coordinates. The information required comprises the position of the sensor given in (X, Y, Z) coordinates in an appropriate reference system, and the attitude of the sensor given as rotations about the reference coordinate system. These six parameters are known as the elements of exterior orientation. They can be found in a number of different ways:

1 By measurement on the platform or sensor. The position can be accurately determined with some satellites, for example ERS-1 where the orbit is accurately determined, or if a GPS receiver is used. There is no suitable system for determining attitude accurately.

2 With the use of GCP. The position and attitude of the sensor can be calculated by relating the image to ground control points. This is relatively straightforward for photographic images, but is more complex when the image is formed over a period of time as with push broom sensors. The sensor geometry and orientation is described mathematically and the unknown parameters determined by a mathematical solution such as a bundle adjustment.

5.5 Principle of correcting height distortion

If the relief is sufficient to cause distortion (section 3.2), then a correction must be applied. The process of distortion due to relief shown in Figure 4 can be reversed to correct for the effect. Figure 8 shows the same geometry as Figure 4 but the terrain is depicted by a series of points, P_1, P_2, ... P_n on a regular grid. Each point has a planimetric position (X, Y) and a height (Z), and together they form a DEM. If a corrected image (orthoimage) is to be formed then the correct plan position of the point P_i , that is P_i', will fall at p_i". To find the correct Gray level value the position p_i in the image is found by projecting P_i into the image. The Gray level value at p_i is placed at p_i" by resampling. In practice this is done as a mathematical 3-dimensional transformation taking into account the position and attitude of the sensor. The process can take place in a digital photogrammetric workstation or using an image processing package. These are discussed in section 2. The topic of digital elevation models is discussed in section 3.

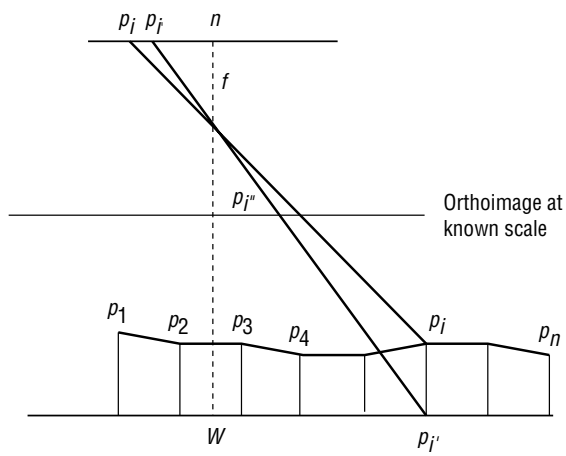


Fig 8. The correction of relief distortion.

6 THE PRODUCTION OF SPATIAL DATA

6.1 Methods

The basic geometry of 3-dimensional imaging systems has been described in sections 3 to 5. This may be incorporated into a computer system in order to allow the extraction of features or height by manual or automatic means. Traditionally, optical mechanical instrumentation has been used by photogrammetrists for map making but this has given way in recent years to analytical instruments in which the operator views the stereoscopic model and follows lines or measures heights under computer control. A degree of automation has also been introduced. Completely digital systems are now being developed in which stereoscopic data acquisition can be carried out as part of a digital image processing system. Such systems allow for stereoscopic viewing and automation is also integral to them.

To use an aerial photograph in a digital system, the photograph first has to be scanned. This is done using a scanner similar to those described in section 2.2, but designed especially for digitising aerial photographs with geometric precision of 1 – 2 μm and pixel size of down to 7.5 μm.

Digital images can also be viewed with a stereoscope on a split screen display or with the anaglyph principle in two colours. Better quality digital display can be obtained by using polaroid or 'flicker' display on the monitor. In these systems glasses must be worn. For a polaroid display the spectacles have complementary polarisation and in the 'flicker' display the glasses are synchronised with the monitor. Stereo viewing systems are available as an option on workstations.

6.2 Digital workstations

Digital photogrammetric workstations (DPWs) have developed from analogue and analytical plotting instruments designed to produce maps from aerial photographs, and from image processing workstations designed for handling satellite data for interpretation and analysis. They are designed to handle stereoscopic data in digital form and will include software which carries out a number of functions automatically. There will normally be a facility for stereoscopic viewing and for automatic matching of images to produce heights. The main functions are:

- Orientation of stereoscopic images
- Feature extraction and assignment of attributes
- Generation of DEMs
- Creation of orthoimages.

Some digital systems have been developed primarily for photogrammetric work, but others have been developed from GIS or image processing systems designed for handling satellite data. Some, such as Intergraph and ERDAS, are closely linked to GIS and data can be easily transferred between systems. There is also a difference in cost and functionality and accuracy of the products from such systems. Table 2 includes the features of some of these systems. A full discussion of digital workstations with references to detailed descriptions can be found in Dowman et al (1992).

One of the major innovations of digital workstations has been the introduction of automatic processing of data, including the formation of digital elevation models by automatic stereo correlation (image matching). The basic principle

Table 2 Some digital photogrammetric systems.

System	*Comment*
Leica HAI-750 Leica HAI-500	Two full photogrammetric systems on Unix or PC platforms. Can use SPOT
Zeiss PHODIS	Full photogrammetric system for use with aerial photographs and satellite data.
Intergraph ImageStation	Photogrammetric data capture as part of digital mapping system. Can use SPOT.
ERDAS Imagine with Orthomax	Designed for production of orthoimages with DEM generation. Linked to ARC/INFO. Can use aerial photographs or SPOT.
R-Wel DMS	PC based system for DEM and feature extraction from images as input to and integration with a GIS.
Leica DVP	PC based system for digital mapping from aerial photographics and SPOT.

underpinning stereo correlation is the matching of small patches from two images: the amount of relative displacement to achieve matching is determined and this is used to determine parallax difference which in turn is used to compute heights. The distortion of the patches because of relief and tilt must be taken into account. Errors will occur because of features such as trees and buildings not on the terrain surface, and matching may be incorrect or impossible if the two images are different because of changes on the ground or large discontinuities in the terrain surface.

This can be a particular problem with SPOT when days, weeks, or months may pass between images being obtained. Accuracies of better than 10 metres can be obtained from stereo matching SPOT data. Line maps have been produced from SPOT using traditional photogrammetric methods. Ordnance Survey in the UK and Institut Géographique National (IGN) in Paris have produced maps at 1:100 000 and 1:50 000 scale using analytical plotting instruments (Smith and Rhind, Chapter 47). Image maps are, however, the more usual product, and can be overlaid with vector data.

Plate 19 shows a stereoscopic pair of aerial photographs in which the buildings appear to lean outwards – thereby demonstrating the effect of relief displacement. This effect is most marked on the left image, which appears in the corner of the photograph from which it was taken. The fiducial marks on the top of the images are used to orient the photographs, and they can be viewed stereoscopically using a pocket stereoscope. Part of Plate 19 shows a plot of part of the area produced on a photogrammetric plotter and an example of the output file for two of its features – a lamppost (marked A), which is stored as a symbol along with its coordinates, and a node on a line (marked B) which is part of a (hatched) building and is attributed a line type (solid) and a colour.

6.3 Digital elevation models

6.3.1 Introduction

DEMs are a way of presenting the elevation of the surface of the Earth in numerical form. A DEM will consist of a series of reference heights arranged in regular or irregular form. A typical regular pattern is a rectangular grid. The spacing of the grid will, together with the accuracy with which the heights are given, be a function of the overall accuracy of the DEM. Digital elevation models can be derived from a number of sources and the processes which are required may differ according to the source. A method of derivation will normally be related to the application of the DEM. The most accurate DEM may give heights to an accuracy of a few centimetres with a high density of reference points, but covering a small area. At the other end of the scale, global DEMs may have a spacing of several kilometres and give heights to an accuracy of hundreds of metres. This section will deal with methods of deriving DEMs from aerial and satellite data. Petrie and Kennie (1990) give a full account of DEMs.

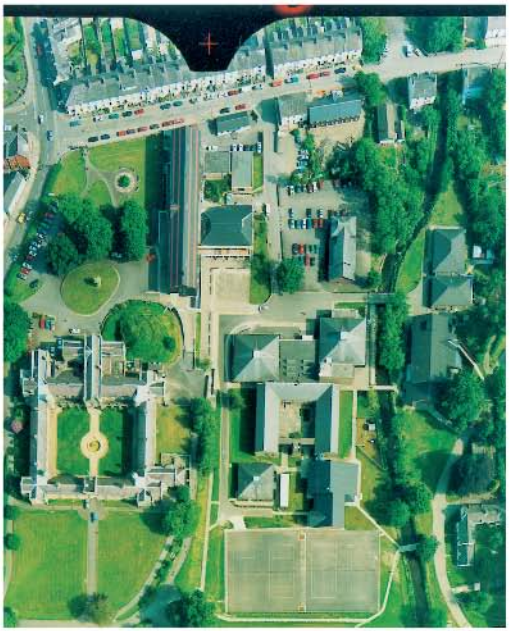
6.3.2 Processes

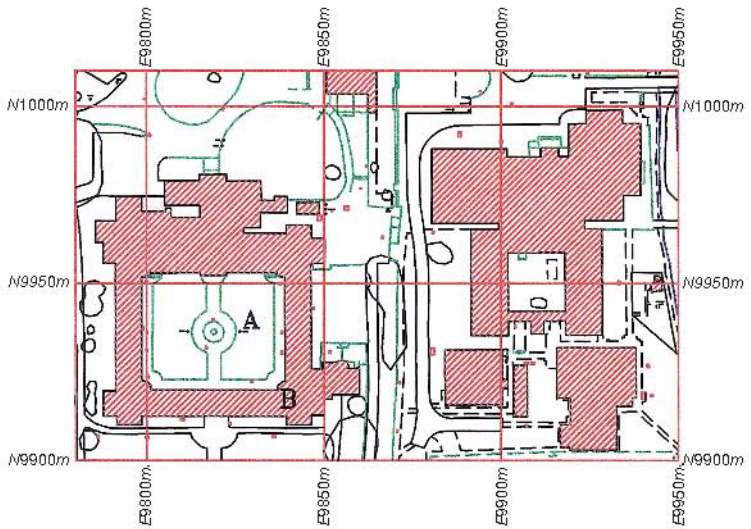
There are three main processes associated with DEMs:

- primary data acquisition;
- resampling to required grid spacing;
- interpolation to extract height of required point.

Each of these stages may introduce errors, the nature of which will depend upon the type of terrain and upon the method used. The data may be acquired from satellites in the form of heights from stereoscopic measurements, from interferometric measurement or from direct measurement with a ranging device. In the case of stereoscopic measurement, automatic correlation will normally be used. In general these methods will produce heights in a quasi-systematic pattern, rows or patches on an approximate grid pattern. These must be resampled to a regular grid. Plate 20 shows a colour coded view of a DEM and a perspective view. The image to the left is a colour coded DEM of the area between Marseilles and Grenoble in southeast France, and is derived from SPOT panchromatic images: the heights were derived on a 50-m grid and are accurate to about 10 metres in height. The image on the right is a shaded view of the same area, derived from the DEM.

DEMs may be generated from stereoscopic aerial photographs or satellite images. The accuracy attainable will depend on the scale of the photograph or on the pixel size. Manually derived DEMs from aerial photography can provide very accurate and reliable information, but the process of recording it is time consuming and tedious. Automatic methods are now widely used but suffer from the problem that a large amount of checking and editing may still be required. The extraction of heights can be done automatically in an off-line mode. Complete scenes of SPOT data can be stereo matched in a few hours

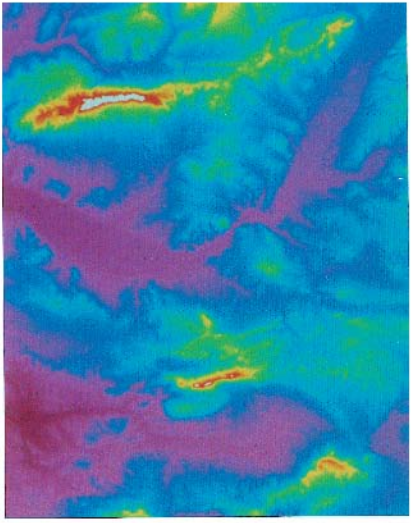




Attribute	Pt A	Comment	Pt B	Comment
Layer	99	Lamppost	122	Building
Mode	3	Symbol	1	Line
Rotation	0.0	Orientation of text		
Width				
Group	18	Post		Solid
Pen	1	Colour	1	Colour
Easting	9826.20		9837.01	
Northing	9936.40		9919.90	
Height	122.66		*130.05	

* The building height is measured on the roof line.

Plate 19
A pair of aerial photographs with a plot and example of the data record.
(*Source*: Cambridge University Collection of Air Photographs)



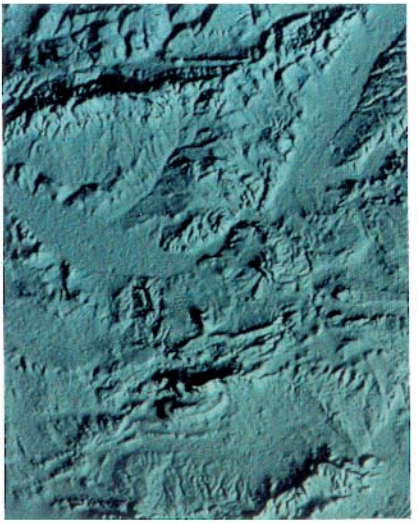


Plate 20
An example of a DEM: (*left*) colour-coded DEM of area of SE France; (*right*) a vertical hill shaded view of same area.
(*Source*: UCL 3D Image Maker)

in order to obtain a DEM which can be used to produce orthoimages and image maps.

In recent years SAR interferometry has been used to create DEMs. This is still an emerging technology but very good results have been reported in some areas. The method has the advantage of being independent of cloud and also that differential methods can detect shifts in terrain, because of earthquake for example, to a few millimetres.

7 ACCURACY OF DATA

Many aspects of data accuracy have been covered indirectly already, and are also discussed by Fisher (Chapter 13) and Heuvelink (Chapter 14). When encoding map data, the principle source of error is the map itself. Encoding systems can reproduce the features on the map with high apparent precision, but the error present in the original document will remain. These errors are attributable to a number of factors:

- original survey;
- revision;
- generalisation;
- drafting and reproduction.

All available information about the source of the map and the data used to compile and revise it should be collected where possible and stored along with the encoded data. Thus, although no reliable measure of the accuracy is available, at least a sensible estimate of the error can be made.

The errors in images will be dependent on the scale of the image, or the pixel size, and on how rigorously the corrections have been applied. A fully rigorous photogrammetric plot will remove all distortions and the errors will be directly related to the scale of the original imagery. An orthoimage produced with a good quality DEM will also have small errors. However, images covering hilly terrain, corrected with a polynomial, may have large errors present. Errors in this type of product will also be subject to the quality of the ground control which may have been poorly identified, or the coordinates might be in error. When registering two datasets, both sets may contribute to the accuracy of the registration.

8 VALIDATION OF ENCODING PROCESSES

Validation can be carried out by evaluating consistency, precision, and accuracy. Accuracy is defined as the closeness of a measurement to the 'truth'. In practice 'truth' usually means the best available data. Accuracy can be evaluated by the root mean square error which is given as:

$$\text{rmse} = \sqrt{\frac{\Sigma v^2}{n}} \tag{9}$$

where v is the residual error, the difference between computed and reference ground coordinates, and n is the number of observations. Precision is a measure of the closeness of measurements to each other and is defined by the standard deviation of a set of measurements:

$$\sigma = \sqrt{\frac{\Sigma(v-\bar{v})^2}{n-1}} \tag{10}$$

where v is the mean of the residuals. For further discussion of RMSE mesures, see Beard and Buttenfield (Chapter 15) and Fisher (Chapter 13).

Consistency is a more subjective measure, depending upon human interpretation of what looks correct.

9 FUTURE PROSPECTS

The growing use of GIS and the increasing volumes of data from new sensors make it increasingly important that data are corrected and referenced to a common system. These corrections need to be as accurate as possible and need in particular to correct for topographical effects. This increases the need for appropriate DEMs. In the future, systems for image processing and image analysis will almost certainly include software to produce orthoimages and DEMs: this latter requirement means that new data sources to produce DEMs are necessary.

A number of missions are planned which are designed for producing stereoscopic data for mapping. The ideal parameters for a mapping satellite are discussed by Light (1990). The US high resolution satellites are the most exciting of these developments since they will produce very high resolution stereoscopic data to produce high precision (x, y, z) coordinates, orthoimages, and DEMs.

Another development of increasing importance is the automation of processes. Automatic DEM generation has been discussed and is a production process. Automatic registration of images is also possible and work on image-to-map registration is in progress. Extraction of some coverage features such as roads is a problem with considerable potential rewards, but is still at the research stage.

References

Dowman I J, Ebner H, Heipke C 1992 Overview of European developments in digital photogrammetric workstations. *Photogrammetric Engineering and Remote Sensing* 58: 51–6

Light D L 1990 Characteristics of remote sensors for mapping and earth science applications. *Photogrammetric Engineering and Remote Sensing* 56: 1613–23

Logan I T 1995 Cost and benefit considerations in data collection and the application of data collection techniques. Paper presented at Cambridge Survey Officers' Conference

Petrie G, Kennie T J M 1990 *Terrain modelling in surveying and civil engineering*. Caithness, Whittles Publishing

Sarjakoski T 1992 *Suitability of the Sharp JX-600 desktop scanner for the digitisation of aerial colour photographs*. International Archives of Photogrammetry and Remote Sensing 29(B2): 79–86

Schreier G (ed.) 1993 *SAR Geocoding: data and systems.* Karlsruhe, Wichmann

Snyder J 1986 *Map projections – a working manual.* Professional Paper 1395. Washington DC, US Geological Survey