2

First principles

2.1 Introduction

The power of GIS, as with other computer programs, can be deceptive: visually impressive but ultimately meaningless results can appear unassailable because of the sophisticated technologies used to produce them (Eiteljorg 2000). The familiar adage 'garbage in, garbage out' is particularly applicable to GIS, and one of our primary aims throughout this book is to provide guidance on how to use this technology in ways to strengthen and extend our understanding of the human past, rather than to obfuscate it. In this chapter we start by providing an overview of the 'first principles' of GIS: the software and hardware requirements, geodetic and cartographic principles, and GIS data models. These provide the conceptual building blocks that are essential for understanding what GIS is, how it works, and what its strengths and limitations are. Although some of these 'first principles' may be familiar to readers who are experienced in cartography and computer graphics, we nevertheless provide a thorough review of each as they yield the foundation on which we build in later chapters.

2.2 The basics

2.2.1 *GIS functionality*

What does a GIS do? Simply providing a definition of GIS and referring to its abilities to capture and manipulate spatial data doesn't provide much insight into its functionality. More informative is to break some of the basic tasks of a GIS into five groups: data acquisition, spatial data management, database management, data visualisation and spatial analysis. Some of the routine tasks performed under these headings are outlined in Fig. 2.1 and described in Box 2.1.

While each of these tasks are important in themselves, above all GIS should be considered as both an *integrated* and as an *integrating* technology that provides a suite of tools that help people interact and understand spatial information. It is important to stress that although the origins of GIS are strongly rooted in digital cartography, GIS is not just about 'maps' nor is it necessarily only about the digital manipulation of the sorts of information and methods that are usually depicted on maps (cf. Longley *et al.* 1999). The use of GIS has a much broader contribution to make in terms of understanding spatial and even space–time relationships between natural and anthropogenic phenomena (Couclelis 1999). Indeed, it is increasingly common to make the distinction between the software tools used to process geospatial data (GIS), and a geographic information science ('GISc')

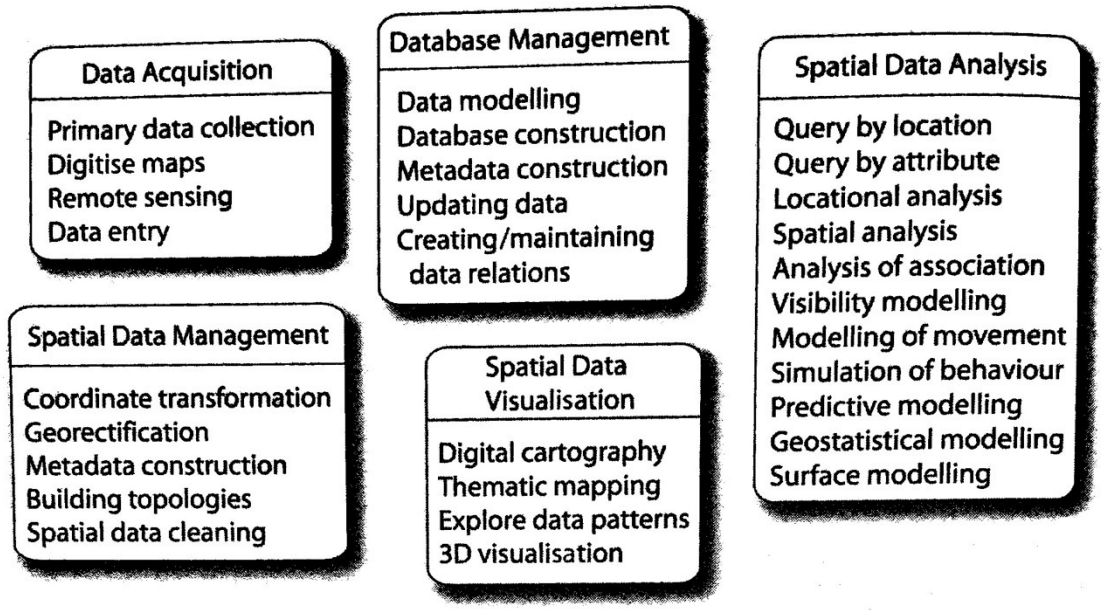

Fig. 2.1 The five main groups of tasks performed by GIS (after Jones 1997, Fig. 1.2).

that is concerned both with the more fundamental conceptual issues of spatial and space–time relationships as well as the impact geospatial technologies are having within the humanities and social sciences (Marble 1990; Curry 1998; Forer and Unwin 1999; Johnston 1999; Longley *et al.* 2005).

2.2.2 *Geographic information*

There is considerable overlap between the aims of the disciplines of archaeology and geography as both share an interest in exploring and interpreting the spatial structure and organisation of human societies at scale from the micro to macro (e.g. Clarke 1977). We thus treat the term 'geographical' as an inclusive one that transcends the discipline of geography (Couclelis 1999). 'Geospatial information' (GI) can therefore be broadly defined as information about natural and anthropogenic phenomena and their relationships with each other. Geospatial information can also describe micro-scale phenomena, such as the patterning of erosion on the facades of historic buildings or the distribution of cut marks on bone (e.g. Marean *et al.* 2001; Abe *et al.* 2002). Most archaeological data – whether artefacts, ecofacts, features, buildings, sites or landscapes – have spatial and aspatial attributes that can be explored using GIS (Fig. 2.2). These attributes include:

- A **spatial location** that tells us where the information is in either a local or global context. A location can be defined by a qualitative term such as 'in Texas' or 'next to the river' or quantitatively using map coordinates. A large component of social-science research uses qualitative positioning data such as counties, cities, census districts or postcodes to form the unit of analysis. Qualitative locations are less commonly used in archaeology

Box 2.1 GIS tasks and descriptions

The acquisition of spatial data GIS is a software platform for the acquisition and integration of spatial datasets. Spatial data include, but are certainly not limited to, topographic maps, site locations and morphology, archaeological plans, artefact distributions, air photography, geophysical data and satellite imagery, all of which can be integrated into a common analytic environment.

Spatial data management GIS uses sophisticated database management systems for the storage and retrieval of spatial data and their attributes. This might involve the transformation of map coordinate systems to enable data collected from different sources to be integrated, the building of vector topologies, the 'cleaning' of newly digitised spatial datasets, and the creation of geospatial metadata.

Database management A major strength of GIS is that it provides an environment for linking and exploring relationships between spatial and non-spatial datasets. For example, given a database on the provenance of a sample of projectile points, and another database that contains information on the morphology of the same points, they can be linked in such a way that it becomes possible to look for spatial patterns in points' morphological variability. Database management, involving conceptual and logical data modelling, is thus an important part of GIS, as is database construction and maintenance to ensure that the spatial and aspatial components of a dataset are properly linked.

Spatial data analysis GIS also provides the ability to undertake locational and spatial analysis of archaeological data, as well as tools for examining visibility (viewsheds) and movement (cost-surfaces) across landscapes. Much work in GIS involves the mathematical combination of spatial datasets in order to produce new data that may provide insight into natural and anthropomorphic phenomena. These range from ecological models that provide predictions of soil suitability for agriculture or erosion potential, or predictive models of potential site location. Tools for geostatistical modelling of spatial data to create, for example, continuous surfaces from a set of discrete observations are also available. GIS can also be a route to the computer simulation of human behaviour and decision making in different types of environments.

Spatial data visualisation GIS has powerful visualisation capabilities used for viewing spatial data in innovative ways (such as thematically or for 'fly-throughs' in three dimensions) that can suggest potential patterns and routes for further analysis. GIS also provide cartographic tools to help produce hard-copy paper maps. Many GIS packages also facilitate the publication of interactive map data on the Internet.

although we may, at times, use locations such as parish, county or survey region. More frequently we use quantitative location data in the form of map coordinates. These include global geographic locational systems, with latitude and longitude being the most common, or national, regional or locally defined Cartesian metric coordinate systems.

- A **morphology** that defines the shape and size of an object, such as 'straight' or '100 m^2'. Qualitative or quantitative descriptors can be recorded as *attribute data* by, for example, recording the size of an archaeological site or the shape of a distribution. Alternatively, it is possible to record spatial morphology directly by mapping the size and shape of a phenomenon, such as an archaeological site on a map. For certain analytical or visual purposes, morphology might be drawn directly on a map, such as the arrangement of a skeleton or the shape of a distribution of artefacts.

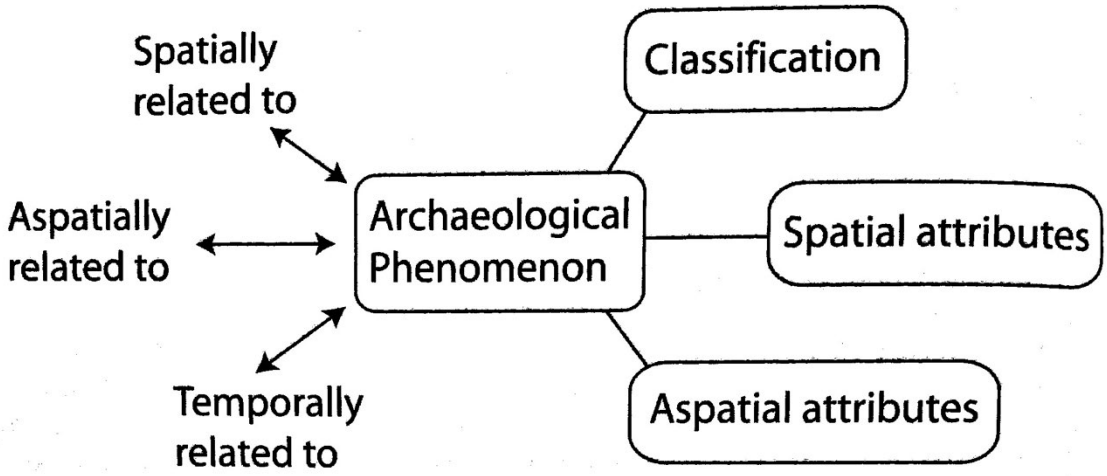

Fig. 2.2 The possible spatial and aspatial characteristics of archaeological data. A burial, for example, is spatially related to other adjacent burials, aspatially related to other burials of that gender or age group, temporally related to other burials of that period, might be classified by the specific position of the body, will have a spatial location and morphology, and possess any number of aspatial attributes (e.g. lists of grave goods and qualitative/quantitative characteristics of the skeleton). After Jones 1997, Fig. 2.8.

- Information about **spatial association and interaction** that describes spatial relationships, such as 'path *a* crosses path *b*', 'from settlement *p* one can see settlement *q*' or 'site *k* is 100 m east of fresh water'. As we discussed in Chapter 1, some types of spatial associations are referred to as *topological*, such as when we talk about path or road connections. Topological relationships are also described as *orientation-independent* because only the connective relationships between objects are important and not their orientation or spatial location. *Orientation-dependent* or *directional* relationships are those that use relational directions, such as above, below, in front, behind, or the cardinal directions east, west, south, north (Jones 1997, p. 25).
- **Temporal relationships** that describe the date and/or associated features in relative terms, like 'contemporary with', 'later than', 'earlier than', etc. Temporal relationships can be important for ensuring that particular types of analysis, such as settlement patterns, are undertaken only on contemporaneous sites.
- One or more **aspatial attributes** that describe the nature of the object. This might consist of a biography of a site or object, information about the colour and raw material of an object, the time of day that a field was fieldwalked, the shape of the cross-section of a feature, or the estimated age of a burial.

The ability to associate aspatial data with spatial objects means that it is possible to explore the spatial characteristics of non-spatial data. For example, given a database of handaxes that records their spatial provenance and aspatial morphological attributes (e.g. weight, size, shape, raw material, reduction stage, etc.) it is possible to explore the relationship between their location and their other characteristics. This is an extremely important ability of GIS that has found application in many areas of archaeological research.

2.2.3 Components of a GIS

A GIS is a computer-dependent technology. In addition to the computer itself, there are a number of other important components to a GIS. The most important ones are:

Software In order to qualify as a GIS the software must have: (i) a spatial database that stores and manages spatial objects; (ii) some mechanism of linking attribute data to these spatial objects, either as an internal function of the GIS package, or by providing functions that enable access to external database systems; (iii) a 'geoprocessing engine', which permits the manipulation and analysis of the geospatial information stored in the spatial and attribute databases. None of the many GIS packages currently available perform all tasks equally well. The choice of software consequently needs to be made with respect to several factors, including the tasks it is needed for, what operating system it has to run under (e.g. UNIX or UNIX-like systems such as MacOS X, Linux, Irix, or Solaris; or one of the versions of Microsoft Windows) and the size of the budget for software, hardware and training costs. A large number of packages are available – too many for us to attempt to review – each with their own strengths and weaknesses in terms of ease of use and the range of analytical tools they offer. An afternoon's research on the web will provide a reasonable grounding in the range of software options. If cost is a primary concern, it is worth knowing that one of the more powerful GIS packages, GRASS GIS,[1] is available free of charge under an open-source licence. Excellent comprehensive commercial GIS packages include Idrisi,[2] the ArcGIS suite of programs,[3] and MapInfo,[4] and all may offer discounts for educational users.

Hardware In addition to the computer that runs the software, which could range from a small palmtop computer to a large institutional mainframe, there are several other hardware components that are essential to making a GIS work. These can be divided into two groups. The first consists of input devices, which might be limited to the keyboard and mouse supplied with the computer, but could extend to digitising tablets, flatbed and roll scanners, digital surveying equipment such as global positioning system (GPS) devices and Total Stations, or geophysical sensors. Chapter 5 discusses the various methods for acquiring digital data in some detail. The second group consists of the output devices needed for viewing and sharing information. A computer monitor is the basic piece of display hardware but, with the obvious exception of the WWW, it is not a very convenient device for distributing information to other people. Some type of printer, from standard letter devices to larger colour plotters, is needed for producing the maps, graphs and tables that GIS routinely produces. We review map production and spatial data communication in Chapter 12.

People GIS operators are the most crucial part of the system as they are responsible for the design and analysis of spatial datasets. A GIS is never a fully objective process – data and questions can rarely be simply 'fed' to a GIS and useful results returned – so it is essential that the specialists responsible for digitising, processing and analysing data are closely integrated with both project design and data collection. This is less of an issue when one researcher is conducting both the project design, data collection and analysis, but in large research projects or commercial archaeological units it is

[1] http://grass.itc.it. [2] www.clarklabs.org.
[3] www.esri.com. [4] www.mapinfo.com.

important to ensure that GIS analysts are included at the earliest stages of the project design to prevent any disjuncture between the envisioned project aims and outcomes. A GIS will rarely contribute in any meaningful way if it is tacked-on as an extra and handed to a 'GIS-person' who has no real understanding of the original goals of the project in question.

2.3 Cartographic principles

2.3.1 *Maps, digital cartography and GIS*

A major element of GIS is the visualisation, management and analysis of spatial data presented in the form of digital maps. It is consequently important to emphasise that all maps, whether paper or digital, simplify the world and present an abstract model of spatial phenomena. Maps can be divided into two basic types:

Topographic maps provide general information about the physical surface of the earth, including natural and human-made features like roads, rivers, settlements and elevation. These exist at a variety of different formats and scales, each suited to particular purposes. Navigational air-charts, for instance, are compiled at a scale which is useful for pilots (1 : 500 000) and emphasise topography, settlements, restricted air-space and airports. In the UK, the Ordnance Survey produce a variety of different topographic maps (in both paper and digital formats) showing elevation, natural and cultural landscape features (including archaeological and historical sites and monuments), roads, towns and villages that are suitable for a range of different applications. The US Geological Survey produce equivalent maps for the USA and most countries have similar organisations (e.g. Canada's Centre for Topographic Information, and Geoscience Australia).

Thematic maps provide specific information about a single feature of the landscape or environment, or display information about a single subject. When the data values vary continuously through space it is common to display them on *isarithmic* maps, which use lines to connect points of constant numeric value, such as elevation (*contours*), temperature (*isotherms*), precipitation (*isohyets*) or even frequencies of hailstorms (*isochalazes*). Other themes are more likely to be displayed on *choropleth* maps, which use shading or symbols to display average values of information in different areas, such as vegetation, geology or numbers of artefacts collected in a survey unit.

To emphasise the differences between traditional paper maps and the dynamical interface that GIS offers it is worth noting some of the constraints of the former (cf. Longley *et al.* 1999, p. 6). Paper maps differ from GIS because they are:

Static The dynamic space-time interactions between objects cannot easily be depicted (e.g. changes in population and settlement patterns, or environmental change). A GIS offers the advantage of enabling exploration of the dynamics of temporal patterning. The University of Sydney Archaeological Computing Lab's TimeMap project[5] is an excellent example of this form of dynamic mapping.

Two-dimensional Multidimensionality cannot be easily depicted on paper. Multivariate spatial data and the three-dimensional representation of topography benefit from multidimensional forms of display available in GIS (e.g. Portugali and Sonis 1991; Couclelis 1999).

[5] `www.timemap.net`.

Flat Representing a curved three-dimensional surface, such as the Earth, in two-dimensions often introduces significant distortion in spatial measurements (see below) and GIS provides facilities for improving this.

Precise The traditional methods of cartographic representation do not allow for the depiction of imprecise, 'fuzzy', boundaries (that occur between, for example, vegetational zones, cultural boundaries, etc.). While this remains a problem for some forms of spatial representation in GIS, there are more possibilities for working with less clearly defined boundaries than paper maps traditionally offer.

Difficult to update Once committed to paper, a map is fixed and can only be updated by producing a new map, whereas a digital map may be updated continuously – even in real-time.

Difficult to relate to non-spatial data The attributes of the objects on traditional maps have to be coded and further information can only be found by reference to a gazetteer. A GIS has several advantages over non-digital systems with regards to attribute data: in particular, a GIS offers more comprehensive data retrieval, ease of update and an ability to explore data patterns more quickly compared with its paper counterpart.

A further major advantage of a GIS over traditional mapping is that a GIS permits the organisation of different components of the same map into different thematic map layers (and thus often referred to as *thematic mapping*), which is the basic way that spatial data are organised within a GIS environment. In practice this means that in one GIS digital display many different elements may be combined, each of which can be individually turned on or off, queried, modified, reclassified and edited. Many analytical functions, such as spatial queries, can operate across one or more layers depending on the need of the GIS analyst. Map layers, or subsets of individual layers, can also be combined to produce new maps at will, providing potential insight into relationships between elements on different themes.

2.3.2 *Map projection systems and geodetic datums*

A basic property of a map is that it is has a spatial context – more properly, *geo-referenced* – by implicitly or explicitly referring to positions on the Earth's surface. Obviously with many maps a precise and absolute spatial context is not important; a quick sketch of the route to a friend's house serves a purpose even though it may be inaccurate and relative. However, when precision and absolute spatial context are important, then an explicit system of measurement is required. As the Earth is a complex shape this is not a trivial process and the science of *geodesy* is concerned with the measurement of the morphology of the Earth's surface. The shape of the Earth is best approximated by a flattened sphere, referred to as either an *ellipsoid* or *geoid*, and positions on it can be defined using *polar* or *geographical coordinates* (Fig. 2.3). *Geographical coordinate systems* define degrees, minutes and seconds north or south of the equator as *latitude* and degrees, minutes and seconds east or west from Greenwich Prime Meridian as *longitude*.

This is an elegant and simple solution for locating positions on the planet. It is less suitable for representing the surface of the Earth on a two-dimensional plane, for example, on a paper map or computer screen. The name given to a system used

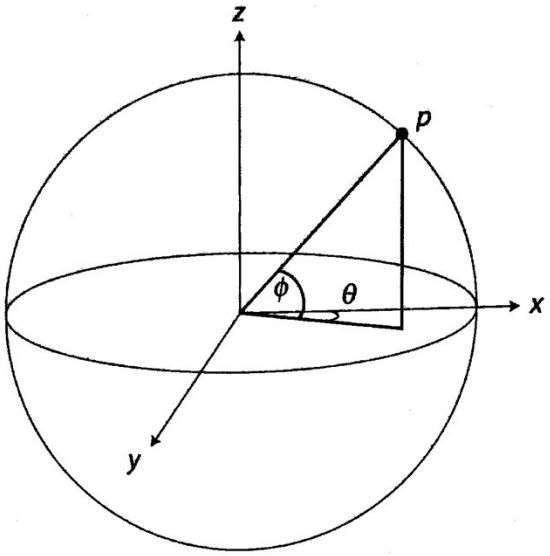

Fig. 2.3 Polar coordinates. The circle of the sphere in the x, y-plane is the equator, and in the x, z-plane it is the meridian. If p is an arbitrary point on the surface of the Earth, then the angle defined by θ is therefore longitude, and the angle defined by ϕ is latitude (after Worboys 1995, p. 143).

to display areas of the Earth's round surface on a flat map is *map projection*, which involves a mathematical *transformation* of the units of longitude and latitude (i.e. *graticules*) to a flat plane. Essentially, a flat map of a large area of the Earth's surface cannot be produced without some form of projection. When mapping areas at the continental or international scale the transformation from three to two dimensions causes profound distortion and spatial error in particular types of measurement. At national and regional scales or larger, the distortion arising from projection to a flat surface causes fewer problems, and national and state mapping agencies have established projections for minimising error within their own boundaries. At very small scales, what we might term subregional or local, the surface of the Earth can be regarded as flat and grid systems can be established and used without reference to geodetic correction. Note here the use of the terms 'large scale' and 'small scale', as this can be a source of confusion. Large scale generally refers to scales of 1 : 50 000 or greater (e.g. 1 : 25 000, 1 : 5000, etc.), and small scale to maps with scales smaller than 1 : 50 000 (e.g. 1 : 100 000, 1 : 1 000 000, etc; Thurston *et al.* 2003, p. 37).

Many forms of map projection have been developed for both global and national mapping purposes and most GIS programs will support many or all of the common ones (GRASS, for example, supports some 123 different projections). Projection systems may be grouped into a *projection family* of which there are three main ones, *conical*, *azimuthal* and *cylindrical*, defined according to how the sphere is projected onto a flat surface (for a mathematical discussion see Iliffe 2000). Each projection family has either a *line of tangency* or two *lines of secancy* that define where the imagined projection surface comes into contact with the Earth, and where there is correspondingly the least distortion (Fig. 2.4). All projections will distort

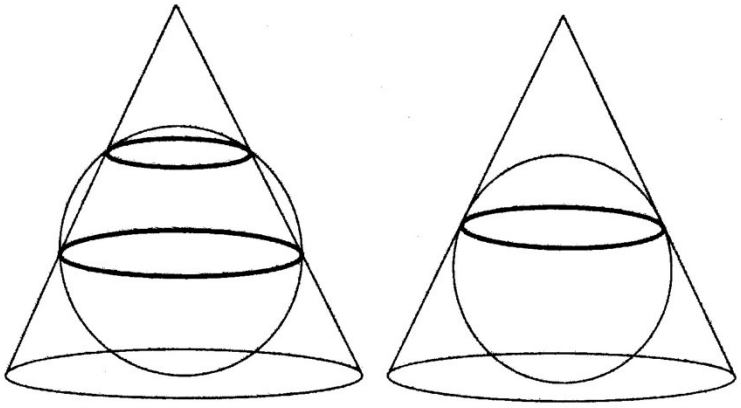

Fig. 2.4 A conical projection with two lines of secancy (left) and one line of tangency (right). The point(s) of contact are also referred to as *standard parallels.*

one or more of the parameters of distance, direction, scale, conformality (shape) and area, although each projection family attempts to minimise distortion in one or two parameters at the expense of increasing it in others.

In addition to the three projection families, there are four projection groups defined on the basis of how this distortion is managed. *Conformal* or *orthomorphic projections* preserve the 90° intersection of lines of latitude and longitude to ensure correct angle measurements between points, but in so doing distort area measurements. *Equal-area projections* preserve area calculations, so that the multiplication of the two edges of rectilinear features represented on a map and globe will be identical (but the properties of shape, angle and scale are then distorted). Projections that maintain distances between one or more pairs of points are described as *equidistant projections*. Any given equidistant projection will only apply to measurements taken in a certain direction: sinusoidal equidistant projections, for example, enforce the measurements parallel to the equator, but distort measurements parallel to the meridian. *True-direction projections* maintain the correct angle from any line measured from the centre of the projection to any other point on the map.

A projection is defined by the combination of a family and then a projection type. For example, a conical projection can be conceptualised as fitting a cone over one of the polar regions as depicted in Fig. 2.4, which is then cut along a meridian as in Fig. 2.5.

The result is a map in which the lines of longitude are straight and convergent, and lines of latitude are concentric arcs. The line of tangency on conic projections is referred to as the *standard parallel* and distortion increases the further one moves away from this line. The amount of distortion can, however, be controlled by altering the spacing of the lines of latitude; if evenly spaced then the projection will be equidistant along the north–south axis (*equidistant conic projection*); if compressed at the northern and southern ends, then the projection becomes equal-area (*Albers equal-area conic projection*).

Azimuthal (or planar) projections represent the Earth's surface on a flat plane using a single point of contact rather than a line of tangency (Fig. 2.6). Azimuthal

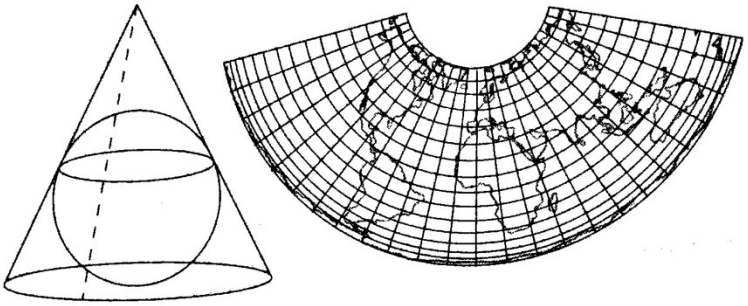

Fig. 2.5 Albers equal-area conical projection with one line of tangency (left) and a meridian (dashed line). The resulting map is to the right, showing the lines of latitude as concentric arcs.

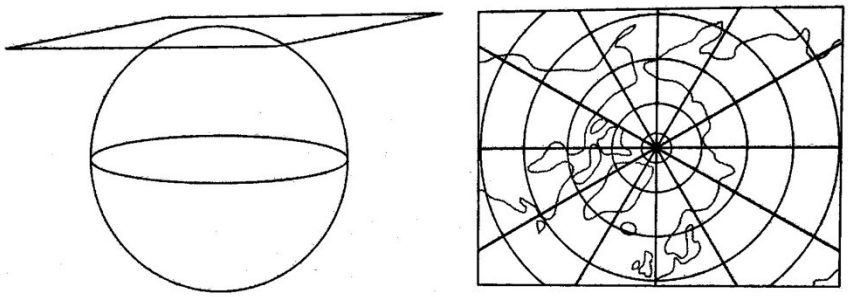

Fig. 2.6 Azimuthal projection with a point of contact at the North Pole. The resulting map has radiating lines of longitude, and concentric lines of latitude. Angle and distance measurements taken along the lines of longitude remain accurate.

(or planar) projections are usually used to map the poles although in theory they can occur anywhere on the Earth's surface. If polar, then the projection is conformal with concentric lines of latitude and radiating lines of longitude. Area distortion occurs as one moves away from the poles, but directions and linear distances from the centre point to any other point on the map are accurate.

Cylindrical projections are conformal and so 90° angles are maintained between the lines of latitude and longitude (Fig. 2.7). Measurements along the line of tangency are equidistant but at further distances from this line area measurements become increasingly distorted.

The most common cylindrical projection is the *Mercator Projection*, which uses the equator as its line of tangency and scales the y-dimension (latitude) to reduce the distortion at polar extremes. This projection gives a very misleading view of the world as movement away from the equator causes areas towards the top and bottom of the map to become disproportionately large in area (Snyder and Voxland 1989, p. 10). The *Transverse Mercator Projection* (TM projection), invented by Johann Lambert (1728–1777), rotates the cone 90° so that a meridian becomes the line of tangency. This distorts measurements in the east–west axis but maintains north–south measurements better than the standard Mercator Projection. The TM Projection is one of the standard ways of mapping the globe.

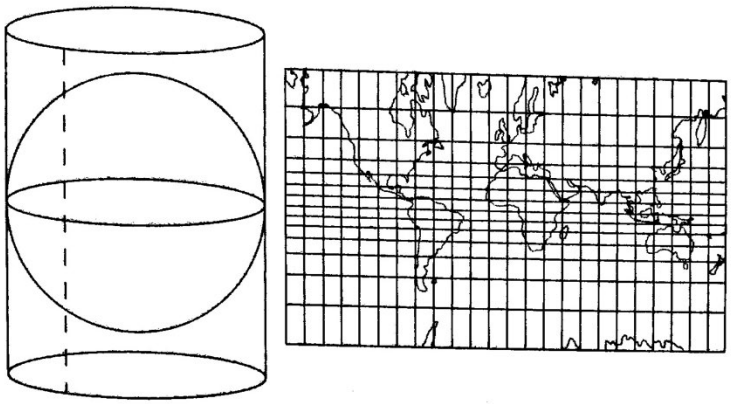

Fig. 2.7 Cylindrical projection with a line of tangency corresponding to the equator and a meridian (dashed line). The resulting map is to the right, showing the lines of latitude as parallel lines.

Finally, the *Universal Transverse Mercator Projection* (UTM) is a twentieth-century modification of the TM Projection that divides the world into 60 vertical zones, each of which are 6° of longitude wide. There is a central meridian in each of these 60 zones that minimises measurement distortion in the east–west to approximately 1 m in every 2500 m (Robinson *et al.* 1995; DeMers 1997, pp. 63–64). Each zone is divided into rows of 8° latitude (12° in the northernmost section) which equates to a 100 000-m wide grid square. The central meridian is given a false easting value of 500 000 m to eliminate the need for negative numbers when specifying east–west coordinates. For the same reason the equator is given a northing value of 0 m for measurements in the northern hemisphere, and 10 000 000 m for measurements in the southern hemisphere. Universal Transverse Mercator coordinates are given by first specifying the zone and then the easting (with 6 digits for 1 m precision) and northing (with 7 digits for 1 m precision). The UTM projection is very popular in GIS and related geospatial technologies like remote sensing because of its global application, minimal distortion and metric coordinate system. Most GPSs are able to record locations in UTM coordinates, making it an ideal system for spatial data collection when a local grid system is not available.

In addition to the projection system used to make the map, it is also important to be aware of which mathematical approximation of the shape of the Earth was used for the construction of a map. The Earth is not an exact ellipsoid, since the surface is not smooth and the poles are not equidistant from the equator. Polar coordinates of latitude and longitude are therefore calculated using a mathematical approximation of the Earth's shape and its centre. Several different approximations have been calculated, often for a specific region of the planet. Clarke's 1866 calculations formed the basis for the 1927 datum of North America (North American Datum 27, or NAD27). NAD27 is being replaced by satellite-derived measurements of an ellipsoid called NAD83 but many organisations still use measurements and locations using the earlier geodetic datum. The Geodetic Reference System

(GRS80), World Geodetic System 84 (WGS84) and European Terrestrial Refer-
ence System (ETRS89) are more recent recalculations of the ellipsoid used in
Europe. National mapping agencies generally use whichever ellipsoid calculations
most closely fit their needs. For example, most national mapping in Great Britain
uses the OSGB36 Datum based on the 1830 *Airy* ellipsoid, although the Ordnance
Survey have adopted the ETRS89 ellipsoid for more recent mapping derived from
GPS receivers.

Coordinate transformation and reprojection

Maps that share a projection system (e.g. UTM) but are based on different ellipsoids (e.g. WGS84 versus NAD27) are not compatible, nor are maps that use different projection systems (e.g. Transverse Mercator versus State Plane) but share the same ellipsoid (e.g. NAD27). For example, the physical distance between two points that have *identical* geographical coordinates, but one based on NAD27 and the other based on NAD83, can be as much as 100 m apart in the USA. For data from multiple map sources to be combined, the maps must share a common projection and geodetic reference system. If this is not the case the projections and/or reference system must be altered through a process called *secondary transformation* or *reprojection*. To compute the transformation, fairly specific information is required about the existing and desired projection systems. Most GIS packages provide tools to transform maps from one geodetic datum to another, and dedicated software tools are also available to help convert between NAD27 and NAD83 (see, for example, the directory of software on the US National Geodetic Survey's website[6] and their online conversion tool[7]). Details about the ellipsoid and projection system used in the construction of a map are typically printed in the corner or are contained in an associated metadata record for digital data (see Chapter 13 for further discussion of metadata elements).

2.3.3 *National and regional grid systems*

Many GIS programs use geographic coordinates of latitude and longitude as the basis for regional maps (most often as *decimal degrees* where minutes and seconds are converted to decimal units so that, for example, 30 minutes 30 seconds is equal to 0.508 of a degree). While decimal degree systems can work well in GIS packages that are able to manage the corrections for spatial measurement, a Cartesian system based on metric units, such as UTM or national (military) grid system, is often a better choice because of the advantages it offers for calculating distances and areas. When a two-dimensional Cartesian grid system is used for mapping, east–west measurements are located on the horizontal x-axis and called *eastings*, and north–south measurements are located on the vertical y-axis and are called *northings* (Fig. 2.8).

At larger scales (e.g. greater than 1 : 1 000 000) most national or regional mapping systems provide metric planar coordinates alongside, or in place of, latitude and longitude. Metric planar coordinates are used in the global UTM projection, the

[6] www.ngs.noaa.gov. [7] www.ngs.noaa.gov/cgi-bin/nadcon.prl.

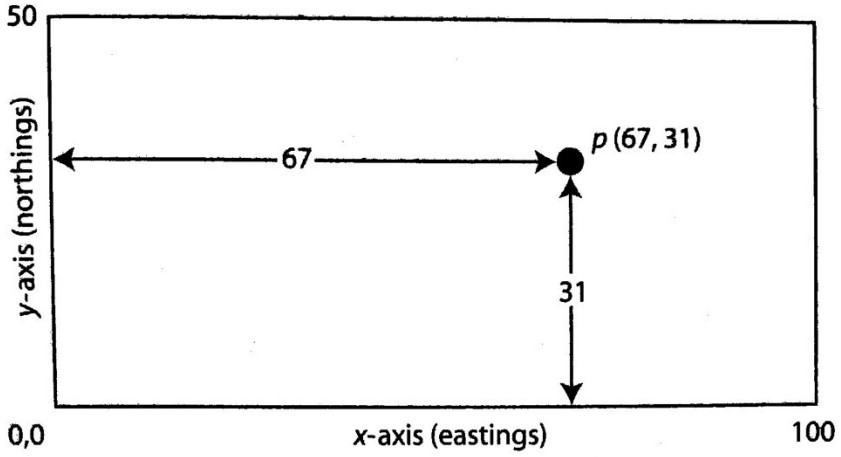

Fig. 2.8 A two-dimensional Cartesian coordinate system. Point p is located by reference to its distance from a 0, 0 datum in the x- and y-planes (respectively referred to as the 'eastings' and 'northings').

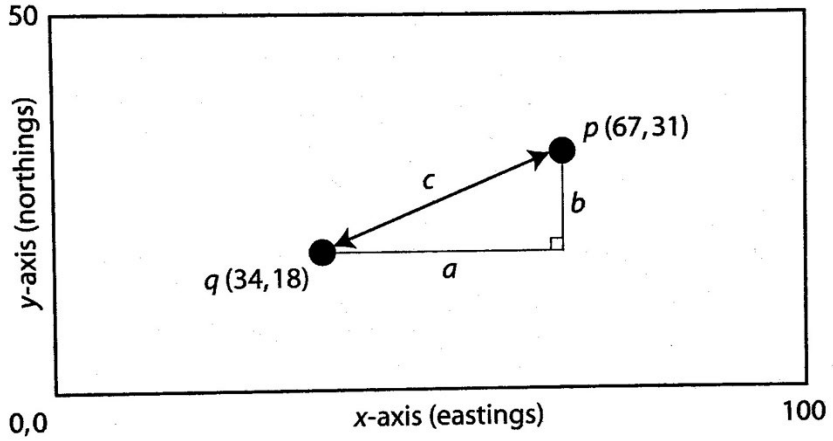

Fig. 2.9 To calculate the linear distance between points p and q (c), Pythagoras' theorem is used: $c = \sqrt{a^2 + b^2}$. As a and b are known ($a = x_p - x_q = 33$, $b = y_p - y_q = 13$), the calculation is $c = \sqrt{33^2 + 13^2} = 35.5$.

US State Plane system, British National Grid and in most other national grids. National grid systems, such as the US State Plane system, are often better choices for regional mapping projects because the ellipsoid is often selected to maximise spatial accuracy for the specific area covered by that particular system. In parts of the world where national or military grids are unavailable, then UTM is an excellent choice. We must emphasise again our warning from the previous section regarding the inevitable and significant spatial errors that will result from combining data derived from maps with different projections and/or ellipsoids.

Metric planar systems have the important and crucial advantage of allowing the easy calculation of distance and area. For example: linear distance measurements can be calculated using *Pythagoras' theorem* (Fig. 2.9); polygon areas can be

calculated using a system that first breaks the shape into smaller *trapezia* and then sums their individual areas to derive the total area; and the geometric centre of a polygon (its *centroid*) can be found by taking the mean of the coordinates of all vertices that define the polygon (for alternatives, see Jones 1997, p. 66; Burrough and McDonnell 1998, p. 63).

2.4 Data models and data structures: the digital representation of spatial phenomena

How does a GIS represent spatial data? The roots of GIS originate with the development of automated mapping in the middle of the last century. In the late 1950s some of the basic computer algorithms for handling geographic information were developed, including the principles for digital cartography, at about the same time that technology had developed to incorporate computer graphics (e.g. Tobler 1959). The *Canada Geographic Information System*, developed in 1963 to manage natural resources, was a natural outcome of these developments and qualifies as the first GIS. It was followed soon after by the development of other systems that were capable of automated mapping (Foresman 1998; Tomlinson 1998). Automated mapping offered considerable time savings over traditional paper methods by providing faster and more accurate facilities for the management and updating of spatial data. These early systems relied on point, line and polygon 'geographic primitives', which still form the building blocks of modern vector-based GIS.

A GIS works by manipulating the digital representations of real world entities. However, a GIS only has a finite set of resources with which to replicate the infinitely complex world and, as a consequence, the digital representations used by GIS are necessarily schematic and generalised. The representation of elements of reality in this way is referred to as a *data model*. In GIS, data models tend to be very simple representations of reality, although as we shall see in later chapters, simple models may become the building blocks for more complex models that are designed to quantify relationships between different entities.

As we saw in Chapter 1, GIS represent spatial data using one or both of the entity and continuous field data models. These are usually implemented as *vector* and *raster* data structures, respectively. Raster and vector data structures store, represent and manipulate spatial information in very different ways. Certain types of entities are more typically represented in one format or the other, although much of the data that archaeologists routinely encounter can ultimately be represented using either structure. Until recently, these two data structures were virtually mutually exclusive: GIS programs tended to rely on either one or the other, forcing users to make a decision as to which they would use. Today most GIS permit the mixing of both raster and vector data as separate thematic map layers, giving users the freedom to decide on the most appropriate structure without necessarily using a different program. The following sections outline the differences between the vector and raster data structures, and provide some examples of ways that different sorts of data are handled by each format.

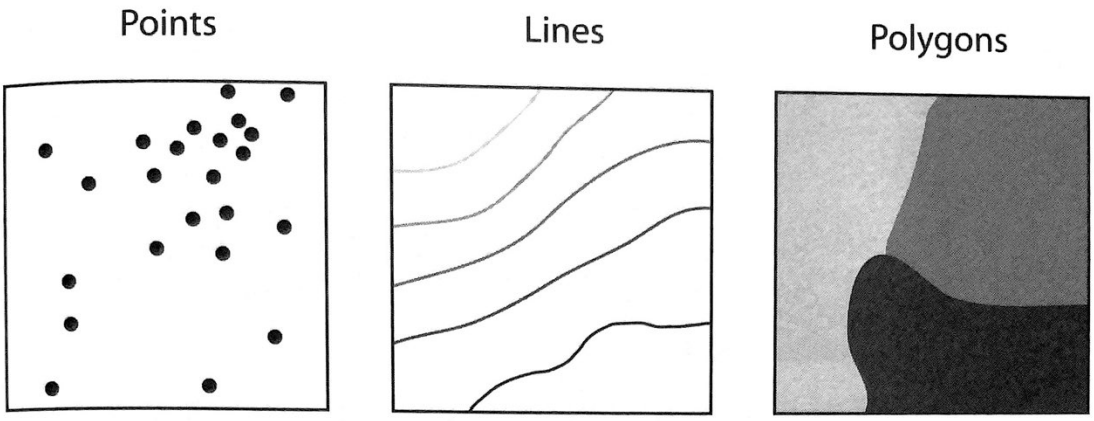

Fig. 2.10 The three vector 'geographic primitives' of points, lines and polygons.

2.4.1 *The vector data structure*

A vector is a mathematical term that refers to one or more coordinates used to define an object in Cartesian space. In the vector structure, real-world entities are represented using one of three geometrical primitives: *points*, *lines* or *polygons*. Each primitive is defined using one or more x, y-coordinate pairs called *vertices*, and are thus described as *discrete* objects because of their precisely defined locations and boundaries (Fig. 2.10). Vertices that are located at the ends of discrete lines, or at their intersections, are called *nodes*.

For example, points are zero-dimensional objects (for they have no length or breadth) defined by a single coordinate pair, and lines and polylines (often also referred to as *arcs* or *edges*) are one-dimensional vectors (having the property of length, but not breadth) defined by two or more coordinate pairs. Polygons, or areas, are two-dimensional objects defined by three or more coordinate pairs. Three-dimensional objects are referred to as *volumes*, but despite the fact that CAD systems routinely use three-dimensional vector objects, the three-dimensional vector structure has not yet been widely implemented in GIS.

The discrete nature of every vector object means that, in addition to possessing its own unique spatial location and morphology, it is a trivial process to provide each vector object with an identification (id) number. On the basis of this unique identifier, each and every object can thus be linked to a set of additional non-spatial attributes that describe additional properties of that object. These properties most often consist of real-world quantitative and/or qualitative variables that give the vector object meaning within the GIS (Fig. 2.11).

Vector topology

An extremely important concept that underlies the vector structure is the geometrical relationships between vector objects, referred to as *topology*. The analysis of topological relationships is explored more fully in Chapter 11, so here it is sufficient to note a few basic concepts. Firstly, topological relationships define

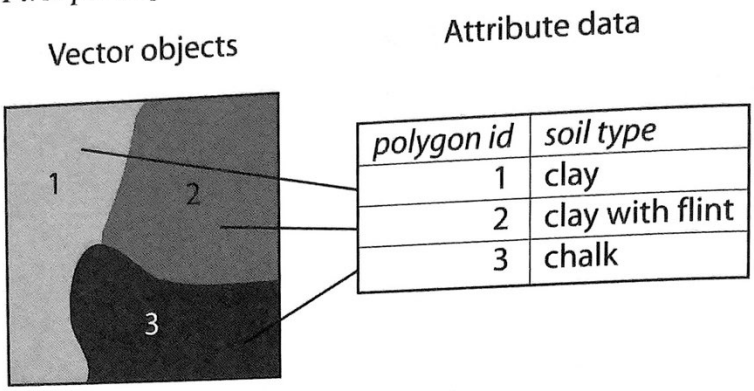

Fig. 2.11 Vector objects linked to attribute data. In this example, each polygon has a unique id number that links it directly to an attribute table that defines the soil type represented by that polygon.

the connections and relationships between vector objects rather than their spatial location. For example, when two roads cross each other, two different topological relationships can potentially exist between those entities. If the lines simply cross without sharing a node, the lines are not topologically connected. This is equivalent to a road crossing another via an underpass and it is not possible to get from one road to another at the point of intersection. If the roads do share a node, they are topologically linked. In this case, it would be equivalent to the two roads meeting at an intersection. Topological relationships are therefore defined by the presence of shared nodes between vector objects. In practice, many GIS require nodes at both crossing and meeting points, in which case additional methods must be used to provide adequate topological information (see Chapter 11).

Topological relationships also define how polygons relate to each other. For example, two adjacent polygons, perhaps representing separate parcels of land or survey zones, are topologically related if they share one or more nodes or arcs in common (Fig. 2.12). Without common nodes this relationship does not exist, and the polygons then must either overlap and/or have a gap between them. It is entirely possible that they intentionally overlap or have a gap to reflect a real-world spatial relationship; but more usually adjacent polygons have an assumed, if not actual, topological relationship. The calculation of spatial relationships and properties of vector objects is not a trivial process, and is dependent both on the data structure and accuracy of the dataset. During the data collection phase and particularly during the process of digitising vector objects, care should to be taken to ensure that topological relationships are properly maintained and defined. Many vector GIS programs have 'clean-up' routines that can be used to create topologies between objects automatically (Chapter 5).

Some geodatabases, such as ArcGIS, provide a set of topological rules to ensure that vector objects are always related in appropriate ways. For example, polygons that define survey areas might have a 'Must Not Overlap' rule, so that any instances

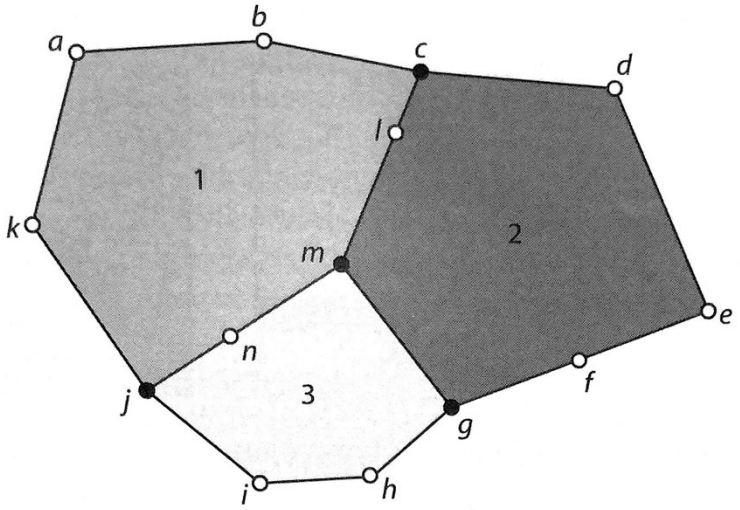

Fig. 2.12 Three topologically related polygons. Polygons 1, 2 and 3 share arcs (edges) defined by nodes *cm*, *mj* and *mg*.

where this occurs are identified and the appropriate action taken (e.g. the overlapping area is subtracted from one polygon, or a new polygon defined by the overlapping area is created).

Topological accuracy also makes for more efficient storage of vector data as vector objects can then share data. Some GIS systems take advantage of this when storing the geometric definitions by only recording an arc (and its vertices) once, and then defining its relationship to polygons. In Fig. 2.12, for example, arcs *cm*, *mj* and *mg* need only be stored once instead of twice for each of the polygon boundaries they define. On large, complex, polygonal maps such as those routinely encountered with soil or geological series, this can result in a significant saving of storage space and computational time, an issue examined in further detail in Chapter 4.

2.4.2 *The raster data structure*

Unlike vector graphics, which use coordinate geometry to define the spatial parameters of objects, raster graphics use a grid matrix of equally sized cells or pixels to represent spatial data (Fig. 2.13). Raster maps are therefore defined only by the number of rows and columns in the grid and the size of each pixel in terms of actual area covered. Each cell also has a value associated with it that represents the attribute status of the object at that location. In a digital elevation model (DEM), for example, each cell has a quantitative value that signifies the mean elevation across the area defined by that pixel, whereas the pixels in a vegetation map may be coded to reflect modal vegetation type. The reliance on pixels and the use of a single attribute per pixel may appear to be very limiting in comparison to the vector structure, but within the simplicity of the raster structure lies its strength. Raster

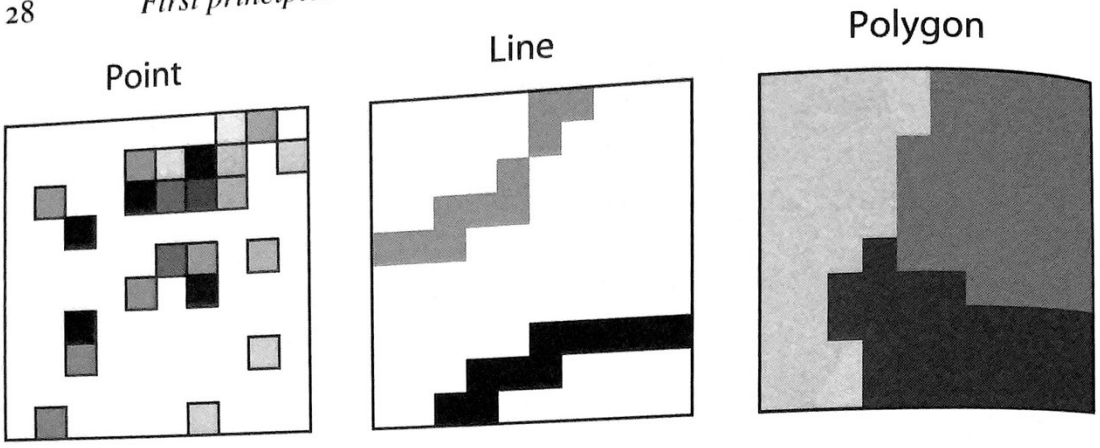

Fig. 2.13 Point, line and polygon primitives as represented on a raster grid.

datasets are easily combined and mathematically manipulated as computers can process and display raster data considerably more quickly than vector data because of the efficiency with which they can store and handle grid data. The simplicity of the structure does not reduce its functionality for, as we shall see in the next chapter, the raster data structure can be used to model some extremely complex spatial phenomena.

A critical variable in the raster structure is the size of the cells, since they define the resolution of a map by providing a minimum unit of representation. Whereas a raster map depicting density of archaeological sites across a large region may consist of a grid of pixels that each represent an area of a square kilometre or more, a raster map showing the density of artefacts across a site may use pixels that define an area of a square metre or less. Although computer processing speed and storage space is continuously increasing, there are nevertheless some practical restrictions on how much information a typical desktop computer can efficiently process. This, of course, varies with the specification of the computer, but loss of performance may be noticed when the total number of pixels is several million or more (e.g. a raster map representing an area of 50×50 km with individual pixel sizes of $10\,m^2$ will require a grid of 5000 columns by 5000 rows and therefore storage of information for 25 000 000 pixels). Decisions relating to resolution need to be made very early on in the model-building process. In the next chapter we show that such decisions can have important consequences for interpretation of the results of analysis and for making sense of spatial patterns.

2.4.3 *Choosing a data structure*

There are many instances where vector systems provide the only sensible means of answering a specific set of questions, or handling the sorts of multiscalar and precisely defined data that one might be interested in exploring. On the other hand, raster data is suitable for powerful spatial modelling, is able to represent continuous datasets more smoothly, and provides image analysis and classification routines suitable for aerial photography and satellite imagery. Many modern GIS systems

are to a large extent 'hybrid' and offer capabilities for manipulating both raster and vector datasets. As a result it is now common to find both data structures being used in a single program environment, thus reducing the choice between the raster versus vector structures to that of appropriateness for the particular needs and questions at hand. From the philosophical perspective adopted in Chapter 1, the vector structure is generally most appropriate when the subject matter has been conceived using the entity model of space. Conversely, the raster structure is usually a better choice if the subject matter has been conceived as a continuous field.

Advantages and disadvantages of the vector structure

Advantages of the vector structure A major advantage of the vector structure is its spatial precision. Real-world entities can be drawn and positioned with an accuracy restricted only by practical limitations such as the precision of the recording equipment. Artefacts, features, sites and other archaeological entities can be integrated in a single environment, each mapped with as much spatial detail as is required for analysis. While centimetre-scale precision would not be required for the analysis of the spatial distribution of sites across a large study region, this level of precision would, however, be essential for the study of the distribution of chert flakes on a knapping floor. Finding the balance between spatial precision and the minimum scale of analysis is crucial: most importantly to prevent spatial errors from influencing pattern recognition. In practice, it is important to recognise that increased resolution also means increased file sizes, with a corresponding burden on storage and processing time. Another significant advantage of the vector structure is that vector objects are maintained as distinct entities and can be easily linked to attribute data records in an internal or external database. For this reason, vector-based GIS programs have traditionally led the way in terms of database integration, as complex attribute queries can be performed with relative ease. A vector map may therefore act as a window into a database, in which each object is described in great detail.

Disadvantages of the vector structure Vector objects are computationally demanding. Every vertex and node of a vector map must be stored in computer memory and drawing vector objects requires a considerable amount of processor time. For this reason, vector data are often much slower to generate on a computer screen than raster data. The manipulation of vector data is correspondingly intensive; spatial queries involving, for example, the calculation of areas of overlap within a large set of polygons needs considerable computer-processing unit (CPU) time. Vector data also impose properties onto real-world objects that do not necessarily correspond with reality. The most important imposition is 'boundedness'. Although many real-world objects do indeed have precise and discrete boundaries, certain types of data are more 'fuzzy' and do not lend themselves to the hard, precise, edges of vector objects. As vector data cannot readily deal with fuzziness or imprecision, this can result in artificial precision in some scenarios. An issue related to fuzzy

boundaries is the implied non-varying state of an attribute across a vector object. For example, a polygon used to represent a discrete survey area may possess an attribute value that represents the density of artefacts in that enumeration unit. This implies a continuous distribution of artefacts in that area, which in reality is rarely the case (e.g. Fig. 12.2). An additional attribute could be used to express the variability within a polygon, but there is no simple way spatially to map continuous change with vector objects. Elevation is therefore inherently difficult to represent using discrete vector objects such as points or lines: contour lines, for example, give an indication of topographic variation at set intervals, but it can be difficult to predict elevation values between the lines. There are special vector structures called *triangulated irregular networks* (TINs) that overcome these difficulties (Chapter 6), but the case remains that some types of data are less well suited to the vector structure.

Advantages and disadvantages of the raster structure

Advantages of the raster structure The speed at which raster data can be processed offers advantages for some applications involving very large datasets and there are several other key areas in which the raster structure can offer advantages over vector formats. Firstly, raster data are very good for mapping continuously varying phenomena, such as elevation, as the continuous cell-based structure is akin to a continuously varying surface. The raster structure is also very good at representing real-world entities that have fuzzy boundaries. For example, a distribution of artefacts collected in a ploughed field could be represented more realistically by using raster cells that show the changing density of material rather than a single polygon that arbitrarily defines the site's area with a single density value. When this type of information is crucial, then the raster data structure offers a clear advantage. Secondly, raster datasets can be mathematically manipulated and combined more easily than polygon maps, making it an exceedingly powerful tool for spatial modelling. A simple model of agricultural potential, for example, may be constructed by combining data from several different sources, such as raster maps of elevation, slope, aspect, soil drainage and soil type in a process called *map algebra*. Thirdly, aerial photographs, satellite images and geophysical surveys produce data in raster formats, and the image processing that is often needed to enhance, classify and make sense of these sorts of data can only be performed in a raster environment.

Disadvantages of the raster structure There are three major disadvantages of the raster structure: its fixed resolution, its difficulty in representing discrete entities and its limited ability to handle multiple attribute data. The first problem arises when data collected at different scales need to be integrated. Combining multiscalar datasets could be seen as introducing additional problems regardless of the data model, and might ideally be avoided, but in practice there are many instances when

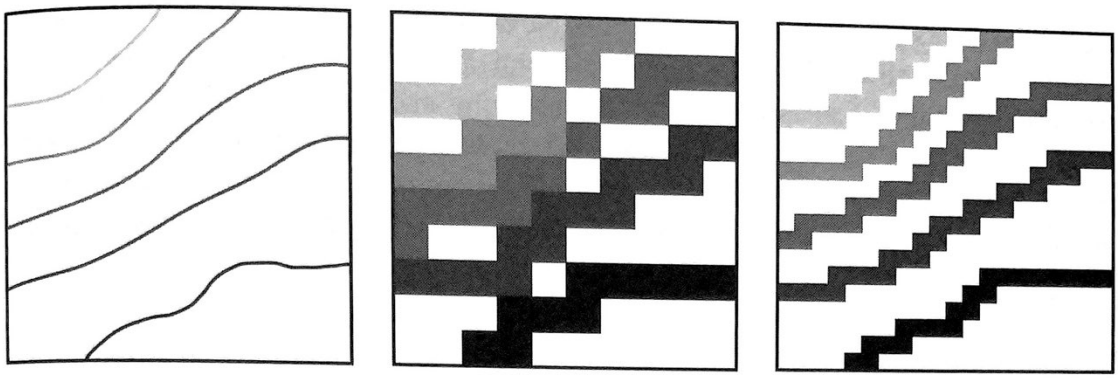

Fig. 2.14 Representing complex curves with raster data can be problematic. The box on the far left shows five vector polylines. The centre box shows the same lines using a 10 × 10 raster grid (i.e. 100 cells). On the far right the resolution has been increasing to a 20 × 20 grid (i.e. 400 cells). This improves the representation, but the raster map still suffers from being blocky and from lost detail.

data collected at different scales must be combined. Field survey data, for example, often mix scales of representation from the larger survey unit (such as a field) to site-based artefact collections where more detail is collected. The representation of multiscalar data is difficult in raster systems and the combination of raster data collected at different scales often results in having to default to the smaller scale and loosing detail. Secondly, problems can arise with representing complex boundaries using raster data because of the inherent limitations of grid data for representing tightly curved objects. Unless the cells are very small in relation to the object being represented and the storage size correspondingly increased, curved lines always will be blocky in appearance (Fig. 2.14).

For this reason complex shapes, such as contour lines, are better modelled using vector objects. Finally, raster data have always been difficult to connect to attribute tables. Although some GIS programs, notably Idrisi and GRASS, provide a facility for linking raster data to a database, in practice this is often more cumbersome than the embedded attribute tables that vector-based GIS programs provide. The raster data structure thus has limitations for the management and querying of multiscalar spatial datasets.

2.5 Conclusion

Geographical information systems (GIS) are a powerful technology that offer a host of analytical possibilities for investigating the spatial organisation of culture and human–environment relationships. These 'first principles' of GIS only define the starting point for exploring the complexity of the human use of space with GIS. In fact, many of these first principles are being constantly challenged by research that is pushing beyond the constraints of two-dimensional mapping to use GIS to model space–time relationships more adequately than the basic vector and raster

building blocks presented in this chapter. Nevertheless, GIS is – for the time being at least – still reliant on cartographic principles and a reductionist tendency that restricts its range of possibilities for representing and interpreting the real world. Within these limitations, however, there is still a very broad range of ways that GIS can be used to develop an understanding of human culture in a spatial framework, and the next chapter provides some real-world examples of how GIS can and does work in archaeology.