Introduction

1

Since their emergence in the 1960s, geographic information systems (GIS) have evolved into an array of sophisticated technologies for working with spatial data. This chapter introduces GIS: what are they and what can one use them for; what technologies and concepts are they based upon; and what makes them different and worthy of special study. Through this Introduction, you will learn to:

- define and describe the main **functions of a GIS** and discuss example **applications** of the technology;
- summarize the key features of the general **computing technologies** that underpin any GIS;
- explain the particular importance of the **database**, which lies at the heart of every GIS; and
- reflect on what makes GIS special, and more generally on why "**spatial is special**."

Sections

What makes GIS special? Most people who work with geographic information systems have asked themselves this question at one time or another. This chapter starts to answer the question by describing the field of GIS against the general background of computing. First, we define the terms "information system" and "GIS," identifying what distinguishes geographic information systems from other information systems. The discussion outlines the main functions of a GIS, with a particular emphasis on those vital functions provided by the *database* (Section 1.1). Then, in Section 1.2, we look at what we can do with a GIS. Presenting some typical, example applications provides a motivation for studying GIS. Spatial data plays a key role in any application of GIS, as do the different ways we use that data to model the world around us, explored in Section 1.3. The fundamental computing technologies and concepts that enable GIS are briefly introduced in Section 1.4. Taking a step back, the chapter concludes with an analysis of what makes spatial—not just GIS—special (Section 1.5).

1.1 What is a GIS?

A good starting point for defining a geographic information system is to look at a general definition of an *information system*. An information system is an association of people, machines, data, and procedures working together to collect, manage, analyze, and distribute information of importance to individuals or organizations. The term "organization" is meant here in the widest sense, to encompass corporations, governments, and societies as well

information system

This chapter has been made available under a CC-BY-NC-ND license.

as diffuse groupings, such as global networks of researchers with common interests or collections of people looking at the environmental impacts of a new development. The World Wide Web (WWW) is an example of an information system. The WWW comprises data (web pages) and machines (web servers and web browsers), but also the many people across the world who use the WWW and the procedures for storing, retrieving, finding, and maintaining information on the WWW.

A GIS is a special type of information system concerned with *geographically* referenced data. Specifically:

geographic information system

> A geographic information system is a computer-based information system that enables the capture and modeling, storage and retrieval, communication and sharing, manipulation and analysis, presentation and exploration of geographically referenced data.

spatial data

Underlying the need for a GIS is the fact that *spatial data*—i.e., data about geographic spaces—requires special handling, management, technologies, and concepts, when compared with other sorts of data (such as found in a library or banking information system, for example). We sometimes also use the term

geospatial

"geospatial" to emphasize the "geographically referenced" aspect of GIS: that data in a GIS is not only spatial, but it is also referenced to the surface of the Earth. Several other related terms and types of information system are relevant to GIS, discussed in Box 1.1 on page 4.

1.1.1 The "shape" of GIS

The world around us is both spatial and temporal, so we have a need for information that has spatial and temporal dimensions. Future decisions that affect us all—for example, in planning new roads or cities, formulating agricultural strategies, and locating mineral extraction sites—rely upon properly collected, managed, distributed, analyzed, and presented spatial and temporal information. A GIS may be thought of as a tool that is able to assist us with these tasks. This tool relies on several underlying functions, summarized schematically in Figure 1.1.

Data storage and retrieval functions At the heart of any GIS is the *database*. A

database

database is a collection of data organized in such a way that a computer can efficiently store and retrieve the data. As we shall see, a database must also be *reliable* (continue to operate even if unexpected events occur, such as power failures); correct and *consistent* (automatically detect and protect data from errors); technology *independent* (with standardized access mechanisms that insulate access from rapidly evolving technological details); and *secure* (ensure sensitive data can be protected, with different access levels for different users). These fundamental capabilities are introduced in Chapter 2. GIS require databases with further specialized capabilities for handling spatial data, which are discussed throughout the book.

Efficient storage of data depends on not only properly structured data in the database, but also optimized indexes and algorithms for retrieval oper-

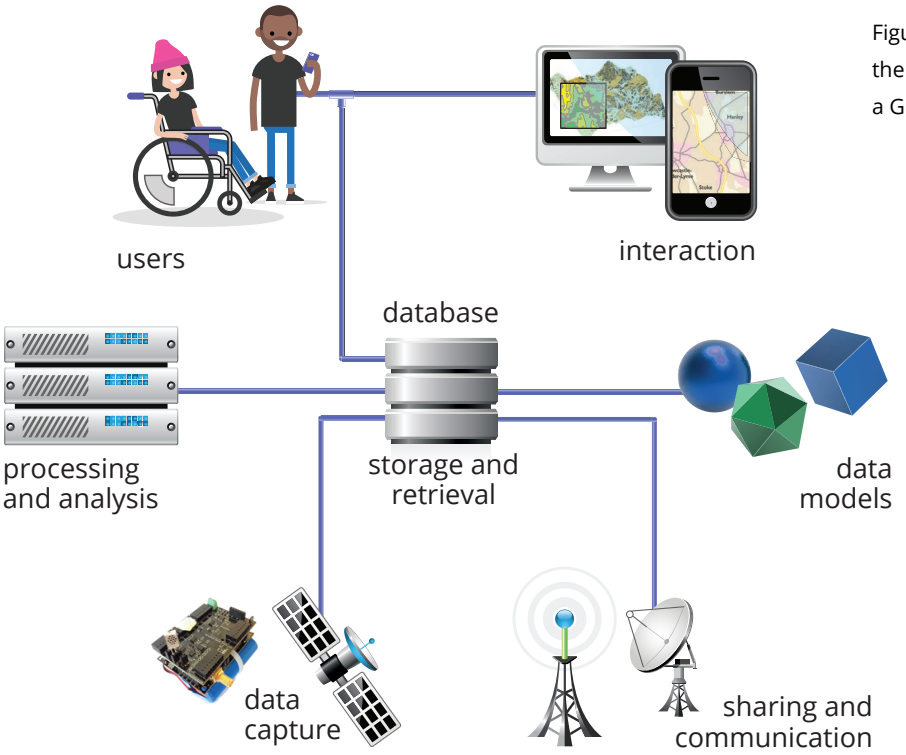


Figure 1.1: Schematic of the functional elements of a GIS

ations. Storage, retrieval, and performance raise many interesting questions for spatial data and will be a further main theme of this text (Chapter 6), also discussed further below in Section 1.4.

Data analysis and processing functions A GIS needs to have sufficiently complete functionality to provide the higher-level analysis and decision support required within an application domain. For example, a GIS that is used for managing utilities and critical infrastructure, such as electricity distribution, water and sewerage, or transportation networks, will require network processing operations (e.g., optimal routing between nodes, connectivity checks, and so forth). A GIS for urban planning will require sophisticated geometry operations and perhaps 3D modeling capabilities.

Identifying and specifying a general set of primitive operations required by a generic GIS is a major concern of the book, covered primarily in Chapter 5. However, the discussion in this book does not address whether specific spatial analyses are suitable for particular applications: those questions we leave to application domain experts.

Data capture functions The process of collecting data from observations of the physical environment is termed *data capture*. The primary source of data for a GIS is from *sensors* that measure some feature of the geographic environment. Surveyors, for example, use sensors to measure distances and

data capture

Box 1.1: GIS terminology

There are several terms in common usage that are closely related to GIS. The term *spatial information system* (SIS) is practically synonymous with GIS. Using "spatial" in place of "geographic" or "geospatial" highlights the broader connections of GIS to concepts and technologies that are not limited to geographic scales, such as room-sized or tabletop spaces. Similarly, a *spatial database* is broadly synonymous with the terms *geographic database* and *geodatabase* and provides the database functionality for a GIS. An *image database* is fundamentally different from a spatial database or geodatabase in that the images have no structural interrelationships, such as *topological* features discussed at length in this book. Image databases use quite different technology to power applications such as medical imagery and image search, although some images may also be geographically referenced (such as a database of satellite imagery). Computer-aided design (CAD) has some elements in common with GIS, and historically some GIS software packages evolved from CAD software. CAD also differs from GIS in that data need not be geographically referenced, and it can instead use coordinate systems local to that design. Closely related to CAD, *building information modeling* (BIM) likewise does not require models to be geographically referenced, and focuses more specifically on the 3D digital representation of physical buildings. The combination of geographic references and structural interrelationships between data means that GIS software is used with larger data sets and/or more complex data models than CAD, BIM, or image databases. Conventionally, and in this book, the term GIS may also be used to refer to "the field of GIS," as well as the information system itself. However, the terms *geographic information science* (GI science), *geospatial science*, or *geoinformatics* are more frequently used, and arguably more appropriate, to describe the systematic study of geographic information and geographic information systems.

angles between features located on the Earth's surface. Satellite imagery uses sensors to measure the electromagnetic radiation reflected or emitted from the Earth's surface at different wavelengths. The field of *Earth observation* (EO) more broadly is concerned with all aspects of sensing data about the Earth, and in particular using space-based remote sensing data sources.

Earth observation

Recent technological advances in MEMS (microelectromechanical systems) have resulted in an explosion in the variety of sensors available today. As you read this book, it is likely that the room you are in or the computing devices around you (perhaps even in your pocket) are bristling with sensors of different types. Miniaturized digital sensors capable of measuring temperature, light, sound, magnetic fields, pressure, and even chemicals are increasingly common in a range of applications and computing devices. For example, *sensor networks* are networks of miniaturized, sensor-enabled computing devices capable of monitoring a wide variety of environmental stimuli, explored further in Chapter 7. Naturally, sensors capable of determining location are of particular importance to GIS, also explored in more detail in Chapter 7.

sensor network

Indeed, technological advances in sensors are being matched by social advances in data capture. An increasingly important source of data is that generated by individual *people* and freely contributed and shared via online repositories, termed *user-generated content* (UGC). Much of this user-generated content is geographically referenced, sometimes also called *volunteer geographic information* (VGI). The motivations for people to capture and share such data are varied, but the impact of geographically referenced user-generated content over recent years has been immense (see Box 1.2 on page 6). As a result, GIS

user-generated content

often need to store, integrate, and process sensor-based information from many different sources.

A secondary data capture stream for GIS is from a legacy data source, such as paper maps. Maps combine the functions of data presentation and data storage; functions that GIS keep separate. Although most maps today are generated from digital spatial data, much of this data was originally captured from paper maps, and many historical maps are still in hard copy format. Converting spatial data stored in a paper map into a form that can be stored in a GIS can be difficult and costly. Automatic conversions, such as scanning a map, cannot easily capture the complex structure of the map, and the results are more similar to an image database than a GIS. Manual conversions, for example, where humans trace the features of a map using a digitizer (Figure 1.2) can capture more structure, but are time-consuming and laborious.

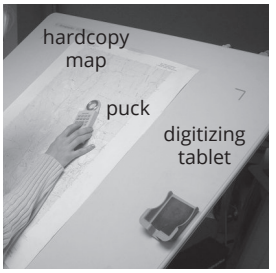


Figure 1.2: Digitizing tablet for manual map digitization

Data sharing functions GIS first began as standalone software in desktop computer labs, and the term GIS is still the label commonly used to identify the software package.[1] Today, however, most applications of GIS rely on the combination of many different software and hardware components, including spatial database servers, web browsers, smartphones and mobile devices, remote sensing satellites and sensor networks, not to mention the different people and organizations that must work together through that application.

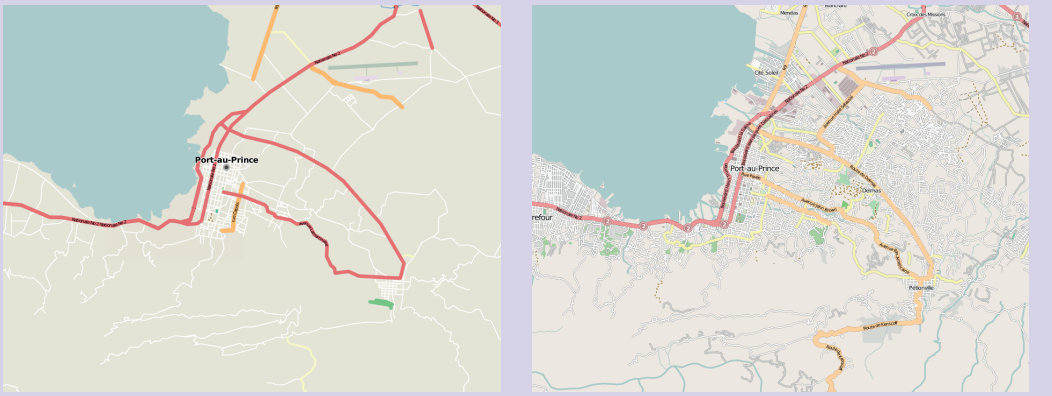
[1] Examples of common GIS software packages today include ArcGIS, QGIS, GRASS GIS, and OpenJUMP GIS.

The different components are frequently not co-located, with data and computation *distributed* across many different physical locations. There are many natural reasons for data and systems to be distributed in this way. Data may be more appropriately associated with one site rather than another, allowing a greater degree of autonomy and easier update and maintenance. For example, details of local weather conditions may be more usefully held at a local site where local control and integrity checks may be maintained. Another advantage of a distributed database is increased reliability; failure at one site will not mean failure of the entire system. Distributed databases may also offer improved performance, sharing the computational load across multiple devices.

Because components may not be physically co-located, most GIS applications rely on a digital communications network (see Section 1.4). For example, in order to access real-time information about the best route to drive home from work, allowing for traffic jams, road works, and changing road conditions, several different distributed information system components must work together with wired and wireless communications networks. Consequently, a fundamental characteristic of any GIS is the capability to *share* data between different information systems, or between different components within a single information system. Understanding and achieving data sharing are key topics in this book, discussed under the general heading of *system architecture* (Chapter 7).

Box 1.2: User-generated content

Perhaps one of the best-known examples of user-generated geographic content is Open Street Map (OSM). OSM combines a conventional spatial database with specially designed software that enables anyone to create and edit topographic data. Founded in 2004, by 2013 OSM contained more than 100 million lines and polygons, made up of more than 1 billion points, generated by more than a million users. In 2022 it had grown to more that 1.5 TB (terabytes) of data. The importance of user generated content first achieved global prominence in January 2010. Within days of a devastating earthquake in Haiti, mappers both in Haiti and around the world had helped to generate previously unavailable detailed and up-to-date maps of the worst-affected areas (see below). These maps were used by many of the humanitarian agencies to aid in their response to the disaster.



12 January 2010 14 January 2010

OSM maps of Port-au-Prince in the days following the 2010 Haiti earthquake (Source: Mikel Maron)

Modeling, interaction, and user functions Data models provide the bridges between the data and analysis functions of an information system and the users and applications of that system. Some information systems only require relatively simple data models. In a library information system, for example, data about books, users, reservations, and loans is structured in a relatively straightforward way. GIS applications, however, demand more complex data models. One of the main issues addressed in this book is the provision of facilities for handling these complex data models. Chapter 2 shows how data modeling is at the core of database design, while Chapters 3 and 4 examine in more detail what makes spatial data and spatial data models special. In preparation, Section 1.3 below introduces some fundamental data modeling concepts.

The bridges that data models provide are critical, because GIS are ultimately as useful as the decisions they support. Effective GIS require the capability to communicate data and analysis effectively to people. Many non-spatial information systems have capabilities for data presentation based around tables of data, numerical computation, and textual commentary. While these forms are also required to support decision-making using a GIS, GIS also require a whole new range of capabilities. These include *cartographic* (map-based) presentation as well as more dynamic, immersive, and exploratory forms of interaction. Interaction with data and analyses results

using a GIS therefore takes us beyond the scope of many traditional software and database systems and is a further major focus of this text (Chapter 8).

1.2 GIS applications

As highlighted above, a GIS is an information system that has some special spatial capabilities. To demonstrate the range of capabilities of a GIS, a series of example applications are described in this section. The examples here are merely to illustrate prototypical applications, rather than in-depth explorations of application areas, however.

The applications have been chosen for a region of England, familiarly called "The Potteries" due to its dominant eponymous industry (Figure 1.3). The Potteries comprise the six pottery towns of Burslem, Fenton, Hanley, Longton, Stoke, and Tunstall, along with the neighboring town of Newcastle-under-Lyme (see Figure 1.4). The Potteries region developed rapidly during the 18th century at the vanguard of the English industrial revolution. The local communities produced ware of the highest standard (for example, from the potteries of Wedgwood and Spode, founded in the late 1700s) from conditions of poverty and cramp. The region's landscape is scarred by the extraction of coal, ironstone, and clay. In the 20th century, The Potteries declined in prosperity although the area is now a focus for regeneration given its historic and geographic centrality in the UK.

Figure 1.3: Historic bottle kilns at the famous Gladstone Pottery Museum in The Potteries today

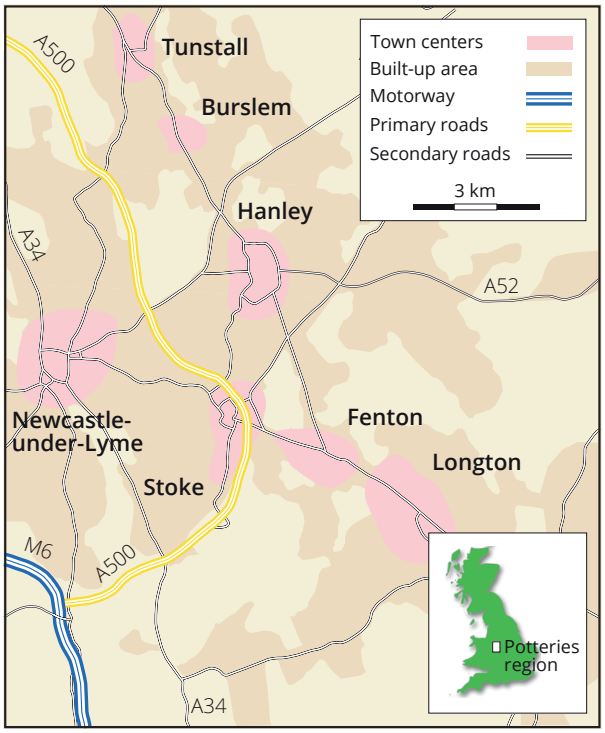


Figure 1.4: The Potteries region

1.2.1 *Resources inventory: A tourist information system*

The Potteries, because of its past, has a locally important tourist industry based upon the industrial heritage of the area. A GIS may be used to support this, by drawing together data on cultural and recreational facilities within the region, and combining this data with details of local transport infrastructure and hotel accommodation. Such an application is an example of a simple *resource inventory*. The power of almost any information system lies in its ability to relate and combine data gathered from disparate sources. This power is increased dramatically when the data is geographically referenced, provided that the different sources of spatial data are made compatible through some common spatial unit or transformation process. Figure 1.5 shows the beginnings of such a system, including some of the local tourist attractions, the major road network, and built-up areas in the region.

resource inventory

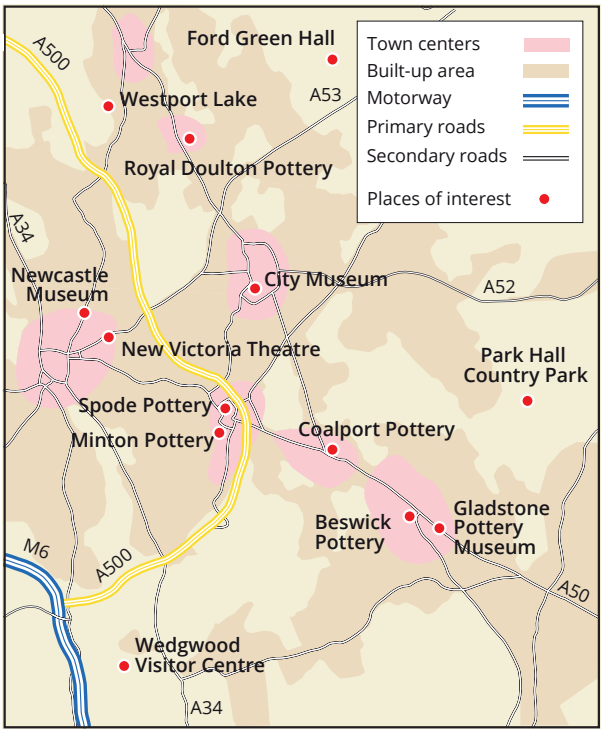


Figure 1.5: Places of interest in The Potteries region

1.2.2 *Network analysis: A tour of The Potteries*

Network analysis is one of the cornerstones of GIS functionality. Applications of network analysis can be found in many areas, from transportation networks to the utilities. As simple example, the major potteries in The Potteries area are famous worldwide. Many of these potteries offer factory tours and have factory outlet stores. The problem is to provide a route using the major road network, visiting each pottery (and the City Museum) only once, while minimizing the traveling time. The data set required is a travel-time network between the potteries: an example is given in Figure 1.6.

network analysis

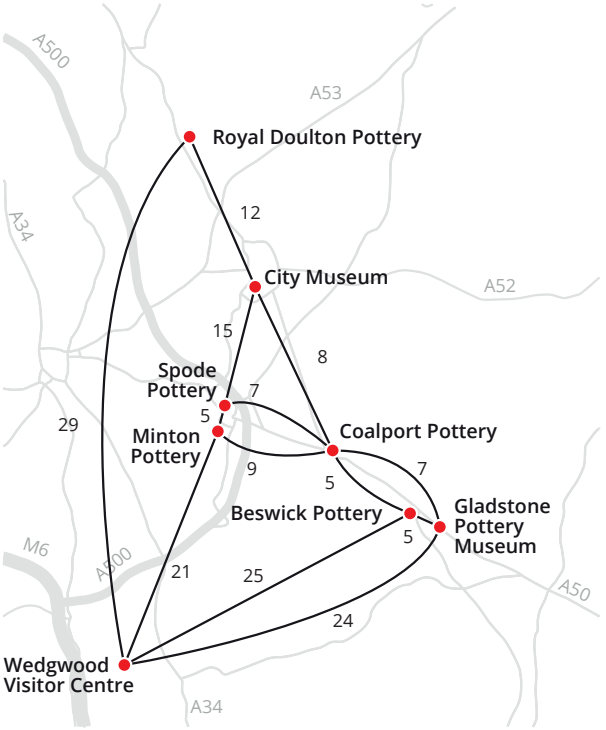


Figure 1.6: Travel network based upon travel times (in minutes)

A travel-time network such as that in Figure 1.6 can be used as a basis for generating efficient routes between the different attractions. The analysis might be dynamic, assigning weights to the edges of the network and calculating optimal routes depending upon changeable road conditions. For instance, one specific relevant analysis is the *traveling salesperson algorithm*, which constructs a minimum weight tour through a network that visits each node at least once.[2]

[2] Can you identify the quickest tour of all the potteries in Figure 1.6? The answer is not straightforward to construct and is discussed further in Chapter 5.

1.2.3 *Distributed data: Navigating around The Potteries*

Much of the data used in a GIS is located in different formats at physically remote locations. A GIS needs to be able to overcome these barriers to data sharing. Continuing the tourist information example, the resources inventory and network analysis above might rely on sharing and analysis of disparate spatial data about the locations of places of interest; transport infrastructure, connections, and congestion; and cultural, hospitality, and recreational resources. These resources are commonly held by different organizations at different locations. Base map data may be held in one place, such as by an online mapping service or national mapping agency. Data about transport infrastructure might be compiled by the local government or held by individual bus or train companies. The tourist information bureau will hold some data about the local amenities, although more will often reside with the individual amenities themselves, such as museums or hotels.

Figure 1.7 shows a schematic version of a distributed information system that might be used as the basis of a Potteries tourist application. Before a tourist visiting The Potteries can receive navigation directions and information about local attractions (for example, on their smartphone), data from all these different sources must be integrated, organized, and processed. Since the tourist will not usually be a GIS expert, these complex tasks will normally be done on behalf of the tourist by some tourist service provider. The service provider might gather all the information needed, either in advance or dynamically when requested, and perform the network analysis necessary to find the best route for a particular user at a particular time and location. Systems that can present data relevant to a user's current location, such as navigation instructions, are called *location-based services*, discussed in more depth in Chapter 7. Although the information actually presented to a tourist at any moment in time may be simple (such as "turn left" indicators, Figure 1.7), the task of integrating data from different sources may be complex.

Figure 1.7: Schematic view of a distributed tourist information system

1.2.4 *Terrain analysis: Siting an opencast coal mine*

aspect

visibility analysis

viewshed

Terrain analysis operates upon data about topographic elevations across a site. Basic information about degree of slope and direction of slope (termed *aspect*) can be derived from such data sets. A more complex type of analysis, termed *visibility analysis*, concerns the visibility between locations and the generation of a *viewshed* (a map of all the points visible from some location).

The applications of terrain analysis are diverse. For our example, the search for new areas of opencast coal mining in The Potteries conurbation has in the past resulted in much interest from local communities, who might be concerned about the effects of such operations. One factor in this complex question is the visual impact of proposed opencast sites. Visibility analysis can

be used to evaluate visual impact, for example, by measuring the size of the local population within a given viewshed. Sites that minimize this population may be considered more desirable.

The terrain surface of the area around Biddulph moor may be represented using a contour map, as in Figure 1.8a. Figure 1.8b shows a perspective projection of the same surface shown in Figure 1.8a. Such projections provide a powerful depiction of the terrain. Figure 1.8c shows the same surface as before, this time draped by the viewshed. The darker shaded regions give the area from which the marked point would not be visible. If we assume the point represents the location of the opencast mine, then the lighter areas provide a first approximation to the visual impact of the mine. Of course, a real case would take into account much more than just the visibility of a single point, but the principle remains.

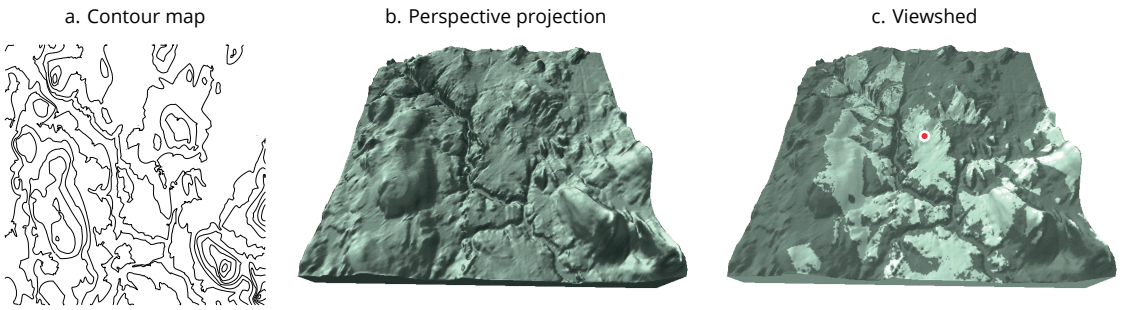


Figure 1.8: Contour map, perspective projection, and viewshed for the marked point on terrain surface

1.2.5 *Layer-based analysis: Extraction sites for mineral ore deposits*

The Potteries area is rich in occurrences of superficial and bedrock sand and gravel, although few such sites have been worked in recent times. Determining the potential of different locations for sand and gravel extraction demands the collation and analysis of data from a variety of sources. Geological data describing the location of appropriate deposits is, of course, needed. Other important considerations are local urban structure (e.g., urban overbuilding), water table level, transportation network, land prices, and land zoning restrictions. Figure 1.9 shows a sample of the available data overlaid on a single sheet, including data on built-up areas, known sand and gravel deposits, and the major road network.

Layer-based analysis results from posing a query such as: "Find all locations that are within 0.5 km of a major road, not in a built-up area, and on a sand/-gravel deposit." Figure 1.10 illustrates the construction of an answer to this question. The shaded areas in Figure 1.10a show the region within 0.5 km of a major road (not including the motorway), termed a *buffer*. The stippled areas in Figure 1.10b indicate known sand and gravel deposits, and in Figure 1.10c the shaded areas indicate locations that are not built up. Figure 1.10d shows the overlay of the three other layers, and thus the areas that satisfy our query. The analysis here is simplistic. A more realistic exercise would take into ac-

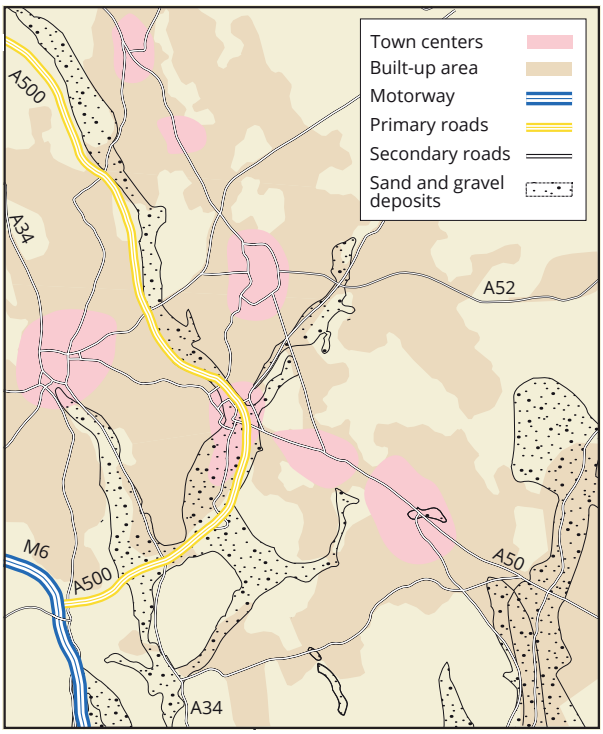


Figure 1.9: Locations of sand and gravel deposits in The Potteries region

count other factors, like the grading of the deposit, land prices, and regional legislation. However, the example does show some of the main functionality required of a GIS engaged in layer-based analysis, including:

buffering

- The formation of areas containing locations within a given range of a given set of features is termed *buffering*. Buffers are commonly circular or rectangular around points, and corridors of constant width about lines and areas.

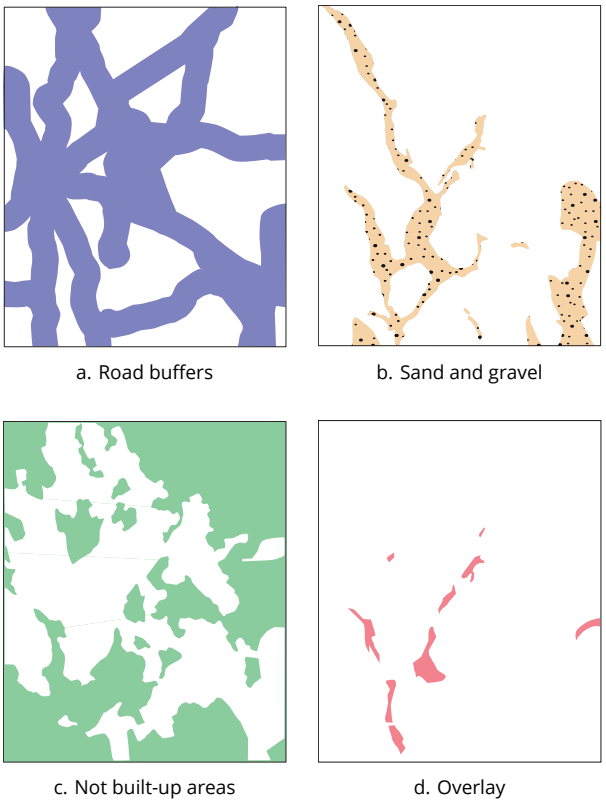
overlay

- The combination of one or more layers into a single layer that is the union, intersection, difference, or other operations applied to the input layers is termed *overlay*.

Layer-based functionality is explored further in the context of field-based models and structures later in the book (Chapter 4).

1.2.6 Location analysis: Locating a clinic in The Potteries

Location problems have been solved in previous examples using terrain models (opencast mine example) and layer-based analysis (estimating the potential of sites for extracting sand and gravel). Our next example is the location of clinics in The Potteries area. A critical factor in the decision to use a particular clinic is the time it takes to travel to it. To assess this, we may construct the "neighborhood" of a clinic, based upon positions of nearby clinics and travel times to the clinic. With this evidence, we can then support decisions to relocate, close, or create a new clinic.



a. Road buffers

b. Sand and gravel

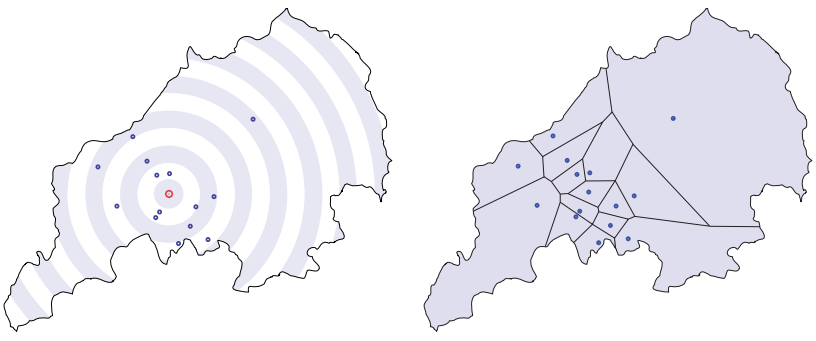
c. Not built-up areas

d. Overlay

Figure 1.10: Layer-based analysis to site a mineral ore extraction facility

Figure 1.11 shows the idealized positions of clinics in The Potteries region. Assuming an "as the crow flies" travel time between points (i.e., the time is directly related to the Euclidean distance between points), Figure 1.11a shows lines connecting locations that are equally far from the clinic in terms of travel time, termed *isochrones*. It is then possible to partition the region into areas, each containing a single clinic, such that each area contains all the points that are nearest (in travel time) to its clinic, termed *proximal polygons* (Figure 1.11b). Of course, we are making a simplistic assumption about travel time. If the road network is accounted for in the travel-time analysis, then the isochrones will no longer be circular and the areal partition will no longer be polygonal. This more general situation is discussed later in the book.

isochrone



a. Clinics, with isochrones around one clinic

b. Proximal polygons

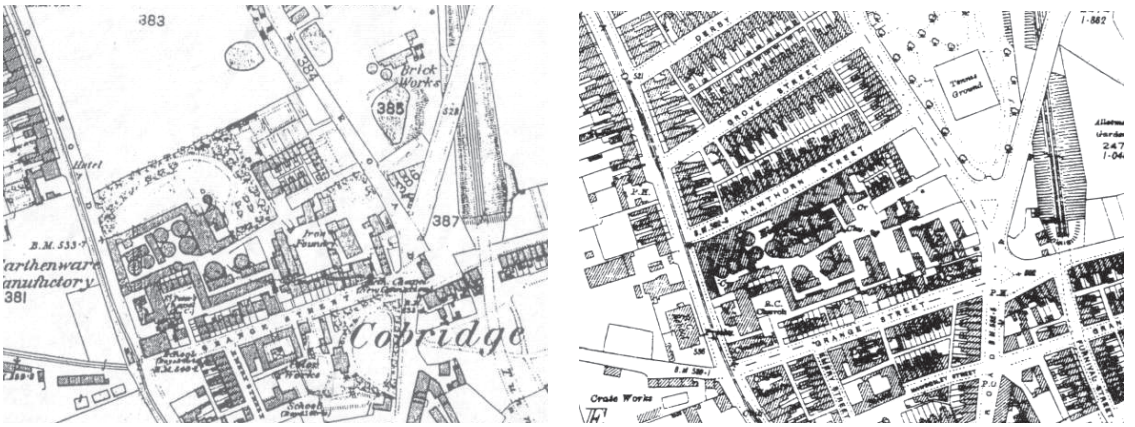
Figure 1.11: Potteries clinics and their proximal polygons

1.2.7 *Spatiotemporal information: Thirty years in The Potteries*

Spatial data sometimes becomes equated with purely static data, thus neglecting the importance of change and time. However, data about the world can always be references to three kinds of dimensions: what (attribute), where (space), *and* when (time). Our next example illustrates the spatiotemporal functionality needed by GIS.

The main period of industrial activity in The Potteries is long since past. The history of the region in the latter half of the 20th century has been one of industrial decline. Figure 1.12 shows the Cobridge area of The Potteries, recorded in snapshot at two times: 1878 and 1924. It is clear that as time has passed many changes have occurred, such as the extension of residential areas in the northwest and southeast of the map. Examples of questions that we may wish to ask of our spatiotemporal system include:

- Which streets have changed name in the period of 1878–1924?
- Which streets have been constructed between 1878 and 1924?
- In what year is the existence of the Cobridge Brick Works last recorded in the system?
- What is the overall spatial pattern of change in this region between 1878 and 1924?



a. 1878 b. 1924

Figure 1.12: History of the Cobridge area, recorded in snapshots at times 1878 and 1924 (Source: Ordnance Survey)

1.2.8 *Digital twins: Smart cities of the future*

GIS as a technology has its origins more than half a century ago. This book (with its own origins more than quarter of a century ago!) has as its focus that growing core of knowledge that lies at the stable foundations of this rapidly advancing technology. Nevertheless, the technology continues to evolve, if anything more rapidly today than ever before. One of the emerging applications of GIS likely to be increasingly important over the coming decade is *digital twins*. A digital twin is a virtual representation of part of our phys-

digital twin

ical world. Many governments around the world are building increasingly sophisticated digital twins that capture not only the appearance of precincts and cities, but are also able to support predictive simulations and fed by near real-time data from sensors, satellites, and citizens (termed *smart cities*).

smart city

As such, digital twins bring together many of the more advanced functions of a GIS in one application. A digital twin of The Potteries does not exist today, at the time of writing, but one is almost certain to be developed in the coming few years, perhaps by the time you read this. Like any digital twin, a Potteries digital twin will need not only 2D mapping capabilities, but 3D representations of the physical topography and buildings (such as in Figure 1.13). It will need to be highly interactive with the ability to support exploration by decision makers and citizens of different views and analyses of the conurbation, populated with frequently updated or even real-time data about the city. It will also need the capability run predictive models, such as of traffic or potential epidemic spread. All this must be supported by rapid storage, update, and retrieval of the spatial data that underpins these capabilities.

Figure 1.13: Digital twins are sophisticated digital representations of parts of our physical world, such as a precinct, city, or nation

1.2.9 Summary of analysis and processing requirements

The examples of applications reviewed in this section demonstrate some of the specialized processing functionality that a GIS needs to provide, usually termed *spatial analysis*. A GIS must handle information about phenomena embedded in the geographic world and having not only multiple spatial and even temporal dimensions, but also structural placement in multidimensional

spatial analysis

geographic models. The key analytical processing requirements that give a GIS its special flavor include:

Geometric, topological, and set-oriented analyses Most if not all geographically referenced phenomena have geometric, topological, or set-oriented properties. Set-oriented properties include membership conditions, relationships between collections of elements, and handling of hierarchies (e.g., administrative areas). Topological operations include adjacency and connectivity relationships. All these properties are key to a GIS and form a main theme of this book.

Field-based analysis Many applications involve spatial *fields*, that is, variations of attributes over a region. The terrain in Section 1.2.4 is a variation of topographic elevation over an area, for example. The gravel and sand deposits and built-up areas of Section 1.2.5 are variations of other attributes. Fields may be discrete (a location is either built up or not) or continuous (e.g., topographic elevation). Fields may be *scalar* (variations of a scalar quantity and represented as a surface) or *vector* (variations of a vector quantity such as wind velocity). Field operations include overlay (Section 1.2.5), slope and aspect analysis, path finding, flow analysis, and viewshed analysis. Fields are discussed further throughout the text.

scalar

vector

Network analysis A network is a configuration of connections between nodes. The maps of most metro systems are in network form, for example, with nodes representing stations and edges representing direct connections between stations. Networks may be directed, where edges are assigned directions (e.g., in a representation of a one-way street system) or labeled, where edges are assigned numeric or non-numeric attributes (e.g., travel-time along a rail link). Network operations include connectivity analysis, path finding (trace-out from a single node and shortest path between two nodes), flow analysis, and proximity tracing (the network equivalent of proximal polygons). Networks are another major topic in this book.

1.3 GIS models and data

Any information system relies on data. However, the data itself is often not the primary focus of the users of an information system; rather, the focus is what that data tells us about the world. This section delves into a little more detail on data, what it is, what are the special characteristics of spatial data, and how we relate data in a GIS to things in the world.

1.3.1 Models and data

The word "model" has been used freely in varying contexts in the preceding text. In general, a *model* can be defined as an artificial construction in which parts of one domain, termed the *source domain*, are represented in another

model

domain, the *target domain*. The purpose of the model is to simplify and abstract away from the complexity of the source domain. Only selected aspects of the source domain are translated by the model into the target domain to be viewed and analyzed in this new target context. Insights, results, computations, or whatever has taken place in the target domain may then be interpreted back into the source domain.

A simple example of a model is a flight simulator. Objects in the real world such as an aircraft, its instrument panel, sounds, movements, views from the cockpit, and the navigation space, are simulated in an artificial environment. The pilot may manipulate the model environment, for example, by simulating a landing into Boston's Logan Airport in bad weather. This experience within the target domain may then be transferred back to experience with flying real aircraft.

The usefulness of a particular model is determined by how faithfully it can simulate the essential elements of the source domain, and how easy it is to move between the two domains. There is a mathematical concept that captures this type of structure, called a *morphism*. A morphism is a function from one domain to another that preserves some of the structure in the translation. Mapping and navigation provide a convenient example. Suppose that the geographic world is the source domain, modeled by a map (target domain). A user needing to travel from Edinburgh to London by road consults and analyzes the map, then translates the results of the analysis back in order to navigate through the UK road network. If the map is a good model of the real road network, then the user's journey may be smooth.

morphism

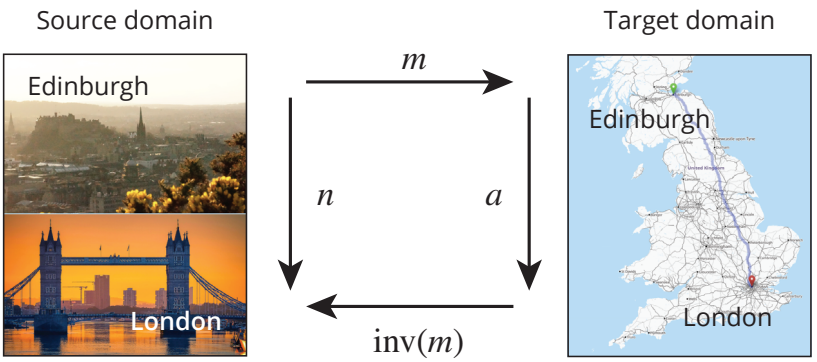


Figure 1.14: Mapping and navigation as a morphism (Map image: Open Street Map)

Figure 1.14 illustrates the concept, with the real-world Edinburgh and London in the source domain, and a mapping engine modeling those locations in the target domain. The figure also shows the modeling process more abstractly in terms of morphisms. The source domain, of real-world cities, is modeled using the morphism m. An activity n in the source domain (e.g., navigating from Edinburgh to London) is modeled by the analysis a in the target domain (e.g., the routing algorithm in the mapping platform). The result of the transformation in the target domain is then reinterpreted in source

domain, using the inverse $\mathbf{inv}(m)$ of the morphism. The whole process "works" if the morphism and analysis accurately reflects the activity in the world.[3]

[3] Mathematically, this is expressed by the equation $\mathbf{inv}(m) \circ a \circ m = n$. Such structural relationships are the subject of the mathematical theory of *categories*. The satisfactory way that the functions (arrows) work together in Figure 1.14 allows us to say that the diagram *commutes*.

In fact, as we take a closer look, we will see that every GIS application encompasses multiple models at different levels of abstraction in parallel. For example, the data in a GIS may embody a model of a particular application domain, such models of transportation or of wildfire. At the same time, the GIS will need to implement certain physical computer-based models that enable efficient storage and retrieval of that data. Indeed, understanding GIS implies an understanding of the different models upon which the technology is based, how these models interact, and their in-built assumptions and limitations.

Conventionally, the different interlocking models in an information system are often decomposed into four distinct models in Figure 1.15. Moving from left to right:

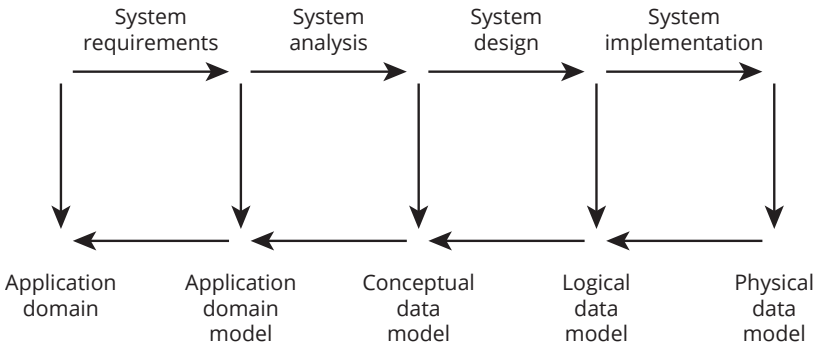


Figure 1.15: Four levels of models in information systems

application domain model

- The application domain itself is the subject of an *application domain model*. An example of an application domain might be bus transportation. In the case of bus transportation, the application domain model might describe entities such as streets and stops, buses and routes, passengers and drivers, as well as more abstract concepts, such as reliability and punctuality. In effect, the application domain model describes and delimits the scope of interest, termed the *system requirements*. Developing an application domain model often requires the assistance of domain experts, enlisted through an initial study of the application domain, sometimes called *requirements analysis*.

system requirements

requirements analysis

- In the *conceptual computational model* or *conceptual data model*, the application domain model is carefully cataloged and structured into computational elements that are more compatible with an information system. Two of the conceptual data modeling techniques we will encounter later in this book are *entity relationship modeling* and *object-oriented modeling*. For example, in the bus transportation example, the entity relationship (conceptual) model might differentiate between which domain elements are entities (such as bus routes and bus stops), which are relationships (such

Box 1.3: System development life cycle

Together, the four modeling functions in Figure 1.15—requirements, analysis, design, and implementation—make up the first four stages of the *system development life cycle*. The system development life cycle is a structured process that can be used to develop any information system, including a GIS application. Two further important stages of the system development life cycle are not modeling tasks. *System validation* follows implementation and aims to validate and refine the developed system by testing it against a range of practical scenarios in deployment. Finally, *system maintenance* is concerned with making ongoing changes and improvements to a system after deployment in response to actual system usage. The logical progression through successive stages of the system development life cycle is often termed the *waterfall model* of system development. In practice, system development is necessarily much more iterative than a steady planned progression from one stage to the next. Issues uncovered at later stages, such as implementation, may throw now light on decisions at earlier stages, such as analysis. Consequently, most system development methodologies today recognize the different stages in the waterfall model, but they allow much more flexible transitions between those stages. For example, *agile development* techniques aim to decompose the development process into multiple rapid development iterations (often called "sprints") that involve cross-functional teams of analysts, designers, developers, as well as users combining in microcosm all the development stages in each iteration.

as "stops at" relationship between bus route and bus stop), and which are attributes (such as route number and stop number). The process of constructing a conceptual model from an application domain model is usually called *system analysis*.

system analysis

- The *logical computational model* or *logical data model* develops the system analysis further by specifying how the conceptual data model can be realized within a specific system paradigm, such as a database or programming paradigm. Several different database paradigms are explored in this book, with most attention paid to relational databases and graph databases. In the case of a relational database, the logical data model will specify the structure of the specific tables (relations) that will be stored in the database, such as tables of bus stops that capture each stop's number or coordinate location. The process of developing a logical model from a conceptual model is called *system design*.

logical data model

system design

- Finally, the *physical computational model* or *physical data model* is constructed by system programmers and developers, who implement the logical computational model for a specific software system or hardware platform. Constructing a physical data model is achieved through a process of *system implementation*, which will involve decisions about exactly how to structure data such that it can be efficiently stored and retrieved.

physical data model

system implementation

The models in Figure 1.15 provide stepping stones to help smooth the transition between application domain and computer system (see also Box 1.3). Without this structure, the system development process quickly becomes unmanageable. Skipping levels tends to lead to systems that are either not useful, because they neglect the human and practical aspects of the application; or inefficient, because they neglect the computational aspects of the system.

1.3.2 *Information and data*

Having pinned down what we mean by the term "model," we should also be precise about the terms *data* and *information*. The structure and interrelationships within data, and how data is collected, processed, used, and understood within an application, form the *context* for data. An understanding of the data model and of the limitations of data are elements of that context. Data only becomes useful, taking on value as information, within this context.

context

data

For example, data about atmospheric conditions is recorded by meteorological stations across the world: there is likely to be one near you recording right now. On its own, the raw data from your nearest recording station is unlikely to be useful to you. Useful information, such as a weather forecast that helps you decide if you need to carry an umbrella today, is produced within the context of careful modeling and analysis of raw data from multiple sources (such as satellite imagery, data from other recording stations, and historical data). Accordingly, information can be defined as "data plus context":

information

$$\text{information} = \text{data} + \text{context}$$

We refine the concepts of data and information in later chapters (particularly Chapter 10), but this basic distinction between data and information is needed throughout the book. Human knowledge about the world and decisions to act in the world are (hopefully) based on information, which in turn is based on data.[4]

[4] Some texts go further and describe a continuum of levels, from data to information, and through knowledge, to wisdom. However, the dividing line between these different levels is frequently fuzzy: one person's knowledge is another person's wisdom. Rather than get tied up in such definitions, we prefer the simpler "information = data + context" in this book.

1.3.3 *Spatial data*

Spatial data is often subdivided into two great classes, *raster* and *vector*. Traditionally, systems have tended to specialize in one or another of these classes, and the division is still evident in many systems today. Figure 1.16 shows a simplified raster and vector representation of the same situation of a house, outbuilding, and pond next to a road.

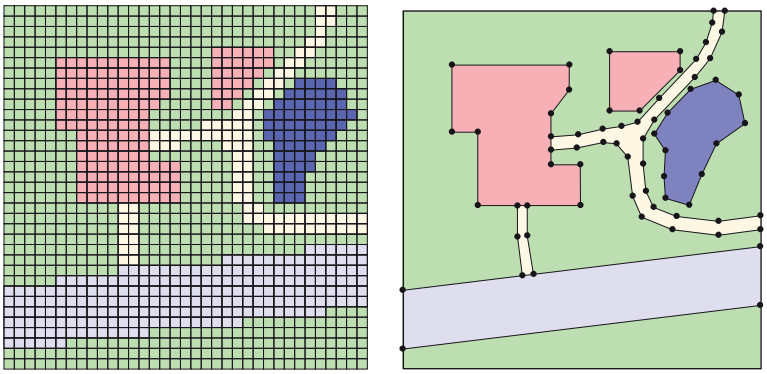


Figure 1.16: Raster and vector data

raster

pixels

Raster data is structured as an array or grid of cells, referred to as *pixels*. Each cell in a raster is addressed by its position in the array (row and column

number). The 3D equivalent of a raster is a 3D array of cubic cells, called *voxels*. Rasters are able to represent a large range of computable spatial objects. Thus, a point may be represented by a single cell, an arc by a sequence of neighboring cells and a connected area by a collection of contiguous cells. Rasters are natural structures to use in computers, because array handling and operations are widely supported by computer systems and languages. However, a raster when stored in a raw state with no compression can be inefficient and unwieldy in terms of usage of computer storage, as the number of cells in the raster increases with level of detail. We will consider efficient computational methods for raster handling in later chapters.

voxels

The other common paradigm for spatial data is the vector format. (This usage of the term "vector" is similar but not identical to "vector" in vector fields). A *vector* is a finite straight-line segment defined by its end-points. The locations of end-points are given with respect to some coordinatization of the plane or higher-dimensional space. The discretization of space into a grid of cells is not explicit as it is with the raster structure. However, it must exist implicitly in some form, because of the discrete nature of computer arithmetic. Vectors are an appropriate representation for a wide range of spatial data. Thus, a point is just given as a coordinate. An arc is discretized as a sequence of straight-line segments, each represented by a vector, and an area is defined in terms of its boundary, represented as a collection of vectors. At comparable levels of detail, the vector data representation is inherently more efficient in its use of computer storage than raster, because only points of interest need be stored. A disadvantage is the vector-representation is not so natural computational a structure to use as arrays, necessitating specialized data storage, retrieval, analysis, and presentation capabilities. We return to these issues in later chapters.

vector

1.3.4 Spatial data retrieval

As identified above, a key function of the database at the heart of a GIS is the capability to efficiently store and retrieve data. To retrieve data from the database, we may apply a filter, usually in the form of a logical expression, or *query*. For example:

query

1. Retrieve names and addresses of all potteries in Staffordshire.
2. Retrieve names and addresses of all employees of Wedgwood Pottery who earn more than half the sum earned by the managing director.
3. Retrieve the mean population of administrative districts in The Potteries area.
4. Retrieve the names of all patients at Stoke City General Hospital who are over the age of 60 and have been admitted on a previous occasion.

Assuming that the first query accesses a national database and that the county in which a mine is located is given as part of its address, then the data may be retrieved by means of a simple look-up and match. For the second

query, each employee's salary would be retrieved along with the salary of the managing director and a simple numerical comparison made. For the third, populations are retrieved and then a numerical calculation is required. For the last, a more complex filter is required, including a check for multiple records for the same person.

There are spatial operators in the above queries. Thus, in the first query, whether a mine is *in* Staffordshire is a spatial question. However, it is likely that the processing for this query needs no special spatial component. In our example, there is no requirement to check whether a mine is within the spatial boundary of the county of Staffordshire, because this information is expected to be given explicitly in the stored address in the database.

In contrast, consider the following three queries (see Figure 1.17):

1. Retrieve the town that contains the lat-long coordinate location **53.0280**, **−2.1747** (a *point query*, answer "Hanley").
2. Retrieve names and addresses of all potteries that lie outside built-up areas (a *region query*, answer "Wedgwood Pottery Centre").
3. Retrieve the closest pottery to Minton Pottery (a *nearest neighbor query*, answer "Spode Pottery").

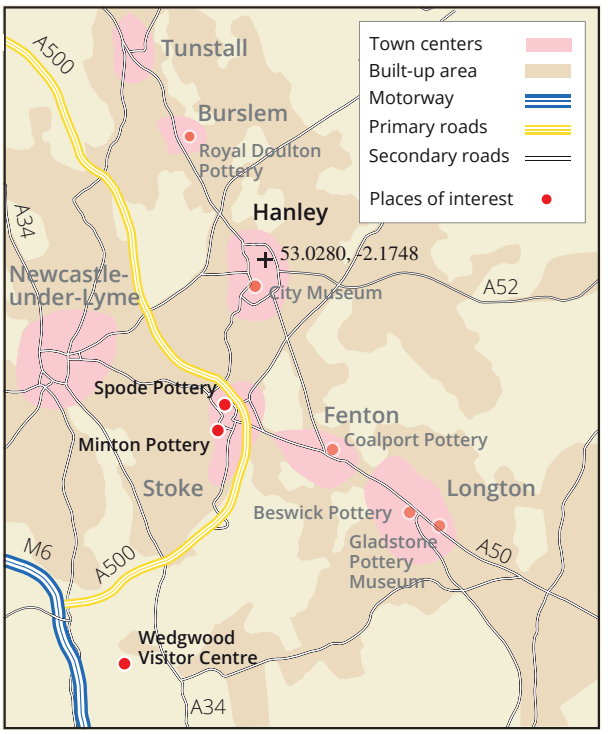


Figure 1.17: Retrieval of spatial data: point, region, and nearest neighbor queries

Responding to these three queries requires specialized spatial data storage and retrieval capabilities. It will not be possible to store the data needed to answer these queries as explicit text or numerical data. Explicitly enumerating every point inside a feature is no more possible than enumerating every region

Box 1.4: Data mining

Data mining refers to the process of discovering valuable information and meaningful patterns within big data sets. A classic example of data mining is *market basket analysis* (MBA). Retailers analyze data about millions of customer purchases to uncover items that are frequently bought together, using that information to promote additional products to customers buying one of the frequent item set. Market basket analysis is often also called *association rule learning*. Other types of data mining include clustering (identifying groups of similar data items), classification (attaching meaningful labels to data items), and anomaly detection (uncovering rare or unusual data items). An example of MBA, often quoted in university lectures, is the story of the discovery of the association between supermarket purchases of beer and diapers by data mining pioneers. The story is memorable because it resonates with deep-seated gender stereotypes about reluctant fathers (as they are presumably the ones purchasing the beer for themselves at the same time as dutifully buying diapers for their baby children, or so the narrative goes). Fortunately, the story is apocryphal, and there is no evidence for such an association occurring in actuality; beer and snacks, such as nuts or chips, is a more veridical market basket association. Indeed, the beer and diapers example serves us better as a cautionary tale about the power of appealing narratives in data mining. As GIS experts, it is advisable to cultivate a healthy skepticism when interpreting the outputs of machine intelligence, as we shall see in Chapter 9.

that contains a feature. Hence, responding to true spatial queries requires new capabilities not available in non-spatial databases.

1.3.5 Spatial data analysis

Now, consider the following two queries:

1. What locations in The Potteries satisfy the following set of conditions:
   - less than the average price for land in the area;
   - within a 15-minute drive of the motorway M6;
   - having suitable geology for gravel extraction; and
   - not subject to planning restrictions?
2. Is there any correlation between:
   - the frequency of vehicle accidents (as recorded in a hospital database); and
   - the locations of designated "accident black spots" for the area?

Satisfying such queries often requires the integration of both spatial and non-spatial information. But it also requires more than spatial data retrieval capabilities; it also requires spatial data *analysis*. Computing the average price for land in an area, the travel time isochrones, and the spatial correlations between data will require us to process spatial data in a way that moves beyond simple point, region, and nearest neighbor queries. While the first of these examples is rather more prescriptive than the second, specifying the type of relationship required, the second example requires a more open-ended, exploratory application of spatial analysis techniques. We will touch on many different types of analysis in this book (and see Box 1.4). However, our focus in this book will be on the computational aspects of spatial data analysis rather than a deeper exploration of the underpinning statistical concepts.

All of the above functionality is attractive and desirable but will be useless if not matched by commensurate performance. Performance is an even bigger issue for a spatial database than a general-purpose database. Spatial data is notoriously voluminous. In addition, spatial data is often hierarchically structured (a point being part of an arc, which in turn forms the boundary of a polygon). We shall see that these structures present problems for traditional database technology. Although databases are designed to handle multidimensional data, the dimensions are assumed independent. However, spatial data is often embedded in the Euclidean plane, requiring special storage structures and access methods. We shall return to questions of spatial data storage, access, and processing later in the book.

1.4 Computing technologies

software
hardware

The term *software* is used to refer to the instructions or programs executed by a computer system. In contrast, *hardware* is used to refer to the physical components of a computer system, such as computer chips, screens, and keyboards. The software needed for a GIS is highly specialized, and much of this book is concerned with those specialized concepts and technologies. However, most of the hardware, networking, and other computing technologies used by a GIS is broadly the same as that used within general-purpose information systems.

This section briefly summarizes the key digital computing technologies used in an information system, to the degree necessary for understanding the role that technology plays in supporting GIS. We assume readers are already somewhat familiar with the architecture of a computer, so this section contains only a brief overview of those components directly relevant to GIS.

1.4.1 The von Neumann model

von Neumann model

Despite the wide range of tasks required of computers, the overwhelming majority of computers we encounter today conform to the *von Neumann model* of computer architecture, developed by the Hungarian-born mathematician John von Neumann at Princeton University during the 1940s and 1950s. According to the von Neumann architecture, a computer system can be thought of as comprising four major subsystems, illustrated in Figure 1.18 and described below.

Processing Data processing consists of operations performed to combine and transform data. Complex data processing functions may be reduced to a small set of primitive operations.

Storage Data is held in storage so that it may be processed. This storage may range from short term (held only long enough for the processing to take place) to long term (held in case of future processing needs).

Control The storage and processing functions must be controlled by the computer, which must manage and allocate resources for the processing, storage, and movement of data.

Input/output Computers must be able to accept data input and to output the results of processing operations. Two important classes of input/output are discussed in more detail in later chapters: input/output between humans and computers and input/output between different computer systems, especially when mediated by a digital communication network.

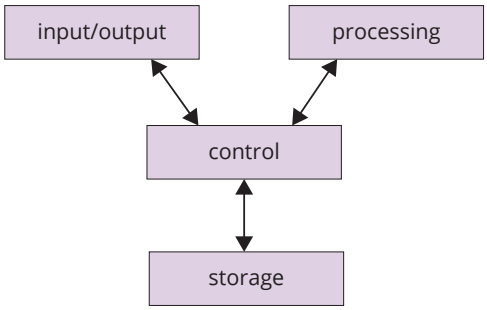


Figure 1.18: The four major functional components of a computer

In terms of components rather than functionality, the key computer system components are the CPU (responsible for processing and control), memory devices (responsible for data storage), and input/output devices (such as human input/output devices and computer networks). Each of these components is considered in turn in the following sections.

1.4.2 *Processing and control*

Processing of data in the computer hardware is handled by the *central processing unit* (CPU). The CPU's main function is to execute machine instructions, each of which performs a primitive computational operation. The CPU executes machine instructions by fetching data into special registers and then performing computer arithmetic upon them.

CPU

The CPU is itself made up of several subcomponents, most important of which are the *arithmetic/logic unit* (ALU) and the *control unit*. The control unit is responsible for the control function, managing and allocating resources. The ALU is responsible for the actual processing functions.

ALU
control unit

Operations are performed upon data sequentially, by retrieving stored data, executing the appropriate operation, and then returning the results to storage. The process of execution is known as the *instruction cycle* (also termed the *machine cycle* or the *fetch-execute cycle*). The instruction cycle involves four steps, shown in Figure 1.19.

instruction cycle

The first step is for the control unit to retrieve an instruction from storage. Next, the control unit decodes the stored instruction to determine what operation must be performed. The control unit then passes this instruction to the ALU, which completes the actual task of executing the operation. Finally,

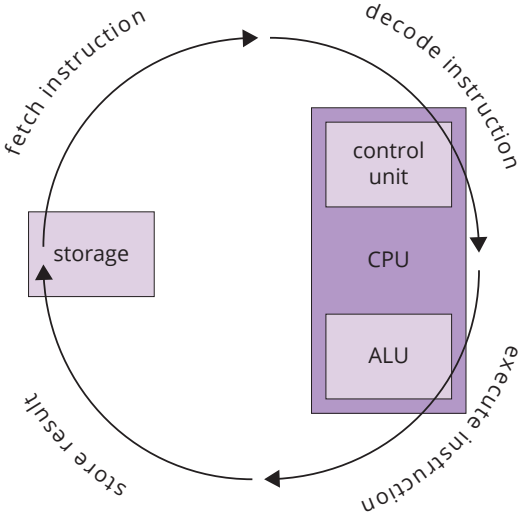


Figure 1.19: The instruction cycle

the results of the execution are returned to storage, ready to be retrieved for a subsequent instruction cycle. Connectivity between the CPU and other components in the computer is provided by dedicated communication wires, each called a *bus*. While almost all CPUs rely on the instruction cycle, computer processors differ in the types and range of instructions they implement (see Box 1.5 on the facing page).

bus

1.4.3 Storage

Digital data must be physically kept somewhere in the computer system. Storage devices differ in their capacity (how much data can be stored), performance (how quickly the data can be accessed), volatility (whether stored data persists after power to the system is turned off), and price. Storage devices can be divided into three categories:

- Storage that can be directly manipulated by the CPU is termed *primary storage*. Primary storage is typically relatively expensive per stored bit, compared to secondary storage, and is generally *volatile* (stored data is lost when the power to the storage device is turned off).
- Storage that can be accessed only indirectly by the CPU (via input/output controllers) is termed *secondary storage*. Secondary storage is relatively cheaper than primary storage and is normally *non-volatile* or *persistent* (stored data persists after the power to the storage device is turned off).
- Storage used to archive huge volumes of data that will be accessed only infrequently by computers is termed *tertiary storage*. Secondary storage is the cheapest of all storage and always non-volatile.

primary storage
volatile
secondary storage
tertiary storage

To enable efficient computation, the speed of access of primary storage must be comparable with CPU instruction cycle times. As a result, primary storage is the fastest, lowest capacity, and most expensive of all memory types.

Box 1.5: CISC and RISC

We usually think of hardware as completely separate from software: hardware is physical computing objects you can touch, software is virtual instructions and data. However, the distinction is not in truth so clear-cut. An illustration of why the distinction between hardware and software can be blurred is provided by "RISC" and "CISC" CPU architectures. The set of instructions supported by a CPU (hardware) is a determining factor in the complexity of programs (software) running on that CPU. In recent decades, most personal computer CPUs have been built using a *complex instruction set computer* (CISC) architecture, such as that found in the Intel x86 family of CPUs dating back to the 1980s. In a CISC architecture, the CPU supports complex instructions, so making software simpler and cheaper to program. However, complex instructions, each of which may require multiple instruction cycles, are generally slower and more energy-intensive to execute than simple instructions. A *reduced instruction set computer* (RISC) architecture supports only simpler instructions than a CISC architecture, each of which can be completed in a single instruction cycle. RISC has the advantage of leading to generally cheaper CPUs with lower power consumption which can as a result run at higher speeds. Many small and mobile computing devices use a low-power RISC architecture, such as the ARM family of CPUs, as well as an increasing variety of laptop, workstation, and supercomputers. However, by making the hardware simpler, RISC architectures typically need more complex software to achieve the same tasks.

Because primary storage is volatile, most information systems, including GIS, also require secondary storage: lower cost, higher volume, non-volatile data storage. Secondary storage is not directly accessible to the CPU. Instead, secondary storage is accessed indirectly, via a bus for transferring data between primary and secondary storage. Secondary storage is consequently much slower than primary storage, with data access times measured in milliseconds or microseconds rather than nanoseconds for primary storage. Therefore, efficient structuring of data files on secondary storage devices is an important factor in GIS performance. Much effort and ingenuity has been spent by computer scientists devising suitable data structures to ensure good performance with spatial data, as we shall see in Chapter 6.

Secondary storage devices can be categorized into disk drives and flash drives. The most important form of disk drive is the *hard disk drive* (HDD). In an HDD, a magnetic disk is coated with a thin layer of magnetic material. The polarization of minute regions of the disk can be accessed or changed by an electrical read/write head. A combination of disk rotation and movement of the read/write head provides access to the entire surface of the disk. *Optical disks* operate along similar principles, instead storing data as minute pits in an optical disk, read by a head that can detect differences in the laser light reflected from the pitted disk surface.

Whether magnetic or optical, disk storage relies on the physical rotation of the disk and movement of read head to store and retrieve data. *Flash storage*, such as solid state drives (SSDs), is based purely on electronic circuitry and has no moving parts. Flash storage relies on the quantum properties of certain semiconducting materials to persistently store data in an array of cells in an integrated circuit. As a result of having no moving parts, flash storage is the fastest form of secondary storage.

flash storage

Finally, tertiary storage is the lowest cost, highest volume, and slowest form of storage, used for archiving and backing up data. Tertiary storage

devices include magnetic tape drives and optical jukeboxes, where libraries of magnetic tapes or optical disks are physically mounted and unmounted by robots. As a consequence, tertiary storage has access times measured in seconds. Figure 1.20 shows some common primary, secondary, and tertiary storage devices.

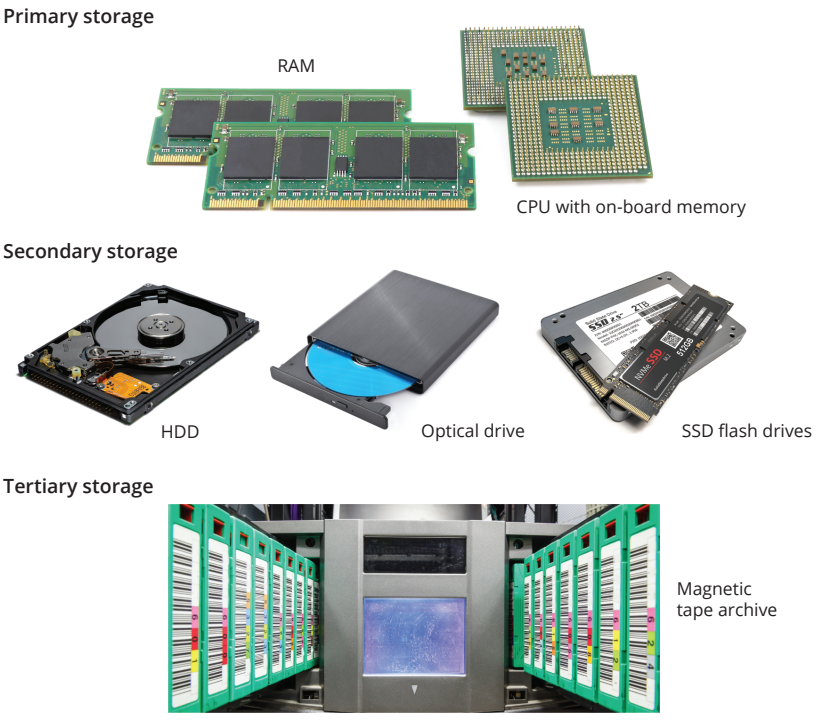


Figure 1.20: Common storage devices

1.4.4 *Input/output and networks*

hard copy
soft copy

Human users are often the focus of input to and output from the computer. For example, *hard copy* output, created by devices such as 2D and 3D printers and plotters, is output with physical permanence. By contrast, *soft copy* output, such as the image on a computer display screen, is transient and intangible. An array of different devices exist to enable humans to input data into the computer and to receive a computer's output, sometimes simultaneously. We return to these devices, their distinguishing characteristics, and their importance to GIS, in the context of GIS interfaces in Chapter 8.

Achieving input and output between one computer and another relies on communication networks, a fundamental component of most of today's information systems, including GIS. Most communication networks are *digital*, meaning a series of binary digits (bits) is transmitted using signal bursts corresponding to the binary values 0 and 1. Digital communication technology has superseded older technology based on *analog* signals, where the signal strength can vary continuously like a sine wave. Figure 1.21 contrasts digital (Figure 1.21a) and analog (Figure 1.21b) signals diagrammatically.

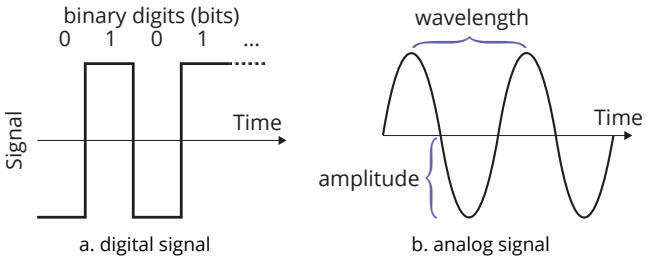


Figure 1.21: Digital and analog signals

Irrespective of whether digital or analog *signals* are being transmitted, all communication networks use electromagnetic (EM) radiation to propagate signals, termed the *carrier wave* or *carrier signal*. EM radiation can be thought of as an analog wave traveling through a medium, such as a cable or the air. The *frequency* of the carrier wave is the number of cycles a wave completes per unit time. At constant wave speed, frequency is inversely proportional to the carrier wave's *wavelength*: the length of each cycle. The *amplitude* is the intensity of the wave.

carrier signal

frequency

amplitude

EM radiation with shorter wavelengths, such as infrared and visible light, can carry more data than radiation with a longer wavelength, such as microwaves and radio waves, because the shorter wavelengths allow the signal bursts to be shorter. However, shorter wavelength EM signals degrade more quickly than longer wavelengths, and so are harder to use over longer distances. The range of wavelengths or frequencies available for data transmission is called the *bandwidth*. There is a clear relationship between the bandwidth of a signal and the amount of data that can be carried by the signal: higher bandwidth means greater data transmission capacity.

bandwidth

Encoding digital data within an analog EM carrier wave involves modulating (varying) either the amplitude or the frequency (or conversely the wavelength) of the carrier signal. A highly simplified example of *amplitude modulation* for encoding digital data within an analog signal is illustrated Figure 1.22.

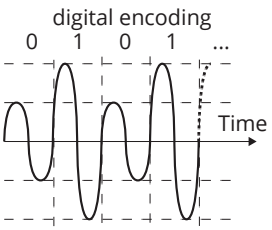


Figure 1.22: Amplitude modulation encoding digital data within an analog signal

Figure 1.23 summarizes the magnetic spectrum and its data-transmission capabilities. Radio waves, microwaves, infrared, and visible light can all be used for computer networks. High-frequency ultraviolet, X-ray, and gamma ray EM radiation are not used for data transmission, because their high energies can be hazardous to the environment and human health.

Signals carried by EM radiation can be transmitted through different media. Ethernet cables and telephone wires are typically made of metal, primarily copper. By contrast, fiber-optic cables are made of fine glass fibers. Although fiber-optic cables are more expensive than copper wire, they offer much higher bandwidth, can operate reliably over much longer distances, are are much less susceptible to interference because they rely on the transmission of visible light when compared with the radio waves transmitted through copper wires.

Data transmission can also take place unguided through the Earth's atmosphere, termed *wireless communication*. Without wires to guide signals,

wireless communication

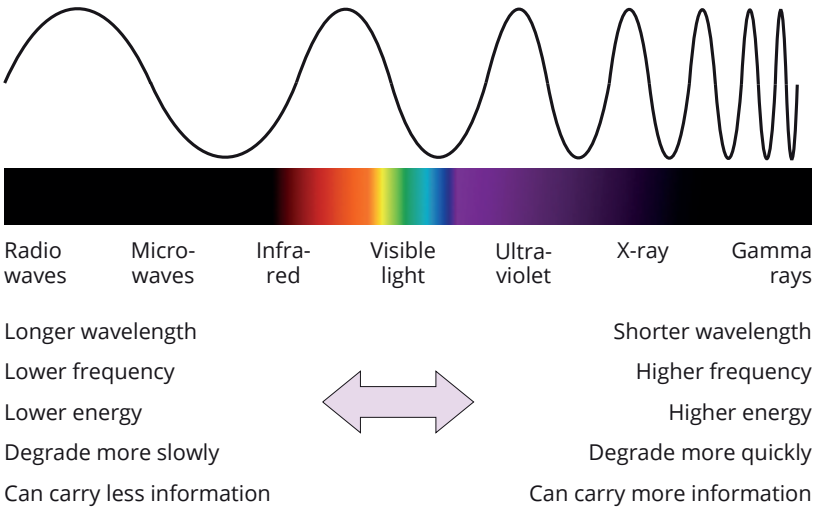


Figure 1.23: The electromagnetic spectrum

wireless communication must overcome signal degradation and interference from environmental radiation (see Box 1.6 on page 32) as well as *path loss*: the reduction in signal strength as it propagates through space. Wireless communications normally use radio wave or microwave EM radiation to transmit data, which carries the advantage that the low-frequency radiation can penetrate non-metallic obstacles, such as walls and floors, while visible light and infrared cannot. As a result, networks based on visible light and infrared signals are often termed *line-of-sight* technology, and are less suitable for wireless communications.

path loss

line-of-sight

Communication networks are also sometimes distinguished according to the size of geographic area they cover. A WAN (wide area network) operates over large-scale geographical regions, such as states and countries. The Internet is the largest WAN in existence, estimated in 2001 to connect more than 100 million hosts, and connecting more than a billion hosts at the time of writing this book.

WAN

A WAN connects groups of LANs (local area networks) together. A LAN connects groups of computers over medium- to small-scale geographical regions, for example, from city centers down to individual buildings. A (wireless) PAN (personal area network) operates over very small geographic or sub geographic areas, typically only a few meters in size. Wireless PANs operate over the shortest distances of any network. A wireless PAN is ideal for connecting multiple small computing devices together, such as smartphones, watches, speakers, personal computers, and sensor networks, without the need for wires. A familiar example of a low-power, short-range radio-wave wireless PAN technology is *Bluetooth*, named after the 10th century Danish king. Figure 1.24 summarizes the different network extents, from a few meters (PAN), through local and metropolitan area networks (LANs and metropolitan area networks, MANs) over a few hundred or thousand meters, to global WANs.

LAN

PAN

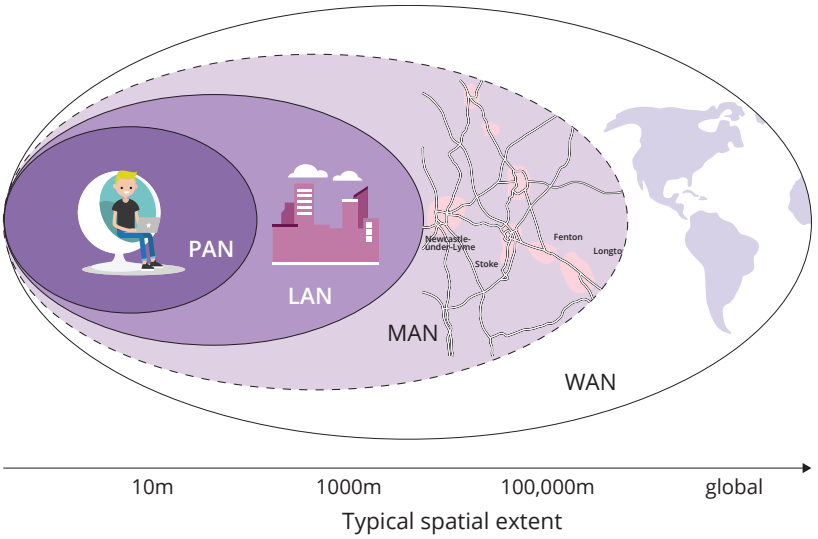


Figure 1.24: Types of network by spatial extent

From a GIS perspective, digital communications have radically altered the way people use spatial information. Digital communications promote sharing of spatial data. Many companies, mapping agencies, and government departments make a variety of spatial data available via the Internet. Digital communications are also *rapid*. The speed of digital communications makes possible certain modes of computing that would not otherwise be possible. For example, mobile GIS applications, such as the tourist navigation system described in Section 1.2.3, rely heavily on rapid and wireless digital communications. We return to the architecture of such systems in Chapter 7.

1.5 What makes spatial special?

The discussion so far has provided many clues to what makes GIS special. We will explore these themes more deeply throughout the remainder of this book. These themes include the following topics we return to later in the book.

- The need already highlighted for specialized techniques for storage and retrieval of spatial data in a database is picked up in Chapter 2.
- Chapter 3 uncovers the richness of different concepts and structures that fall under the heading of "spatial" and need to be supported by a spatial database, including geometry and topology.
- This richness and diversity in turn gives rise to the variety of different models encountered in connection with spatial data in Chapter 4.
- Armed with these concepts, Chapter 5 explores the foundations of algorithms for efficient computation with spatial data and the solutions to iconic spatial problems, such as computing shortest paths, point-in-polygon, and polygon overlay.

Box 1.6: Combating wireless interference

The problem facing all wireless network technologies is how to ensure that signals from different devices do not interfere with one another, causing data loss. There are essentially three increasingly sophisticated mechanisms for achieving this. The simplest option is to ensure that each device only uses a narrow frequency range for communication. This is rather like the way radio stations broadcast only on particular frequencies, into which you need to tune to hear the broadcast. The second option is to ensure that each device only transmits for a certain amount of time on its frequency range. This allows the same frequency to be shared by several devices, rather like the way that different radio programs occur at different times on the same radio station. The third option is to ensure that each device transmits at a range of frequencies, hopping between frequencies at different times in some sequence agreed by transmitter and receiver, termed *spread spectrum* technology. This would be rather like starting to listen to your favorite radio program on one station, and part-way through retuning your radio to another station for the remainder of the program. This "frequency hopping" makes the network more tolerant to interference from other devices and environmental background noise: if some of the signal is lost to interference, the chances are that the signal will be regained next time it hops to a new frequency. Spread spectrum technology commonly switches hundreds of times every second between dozens of randomly chosen frequencies.

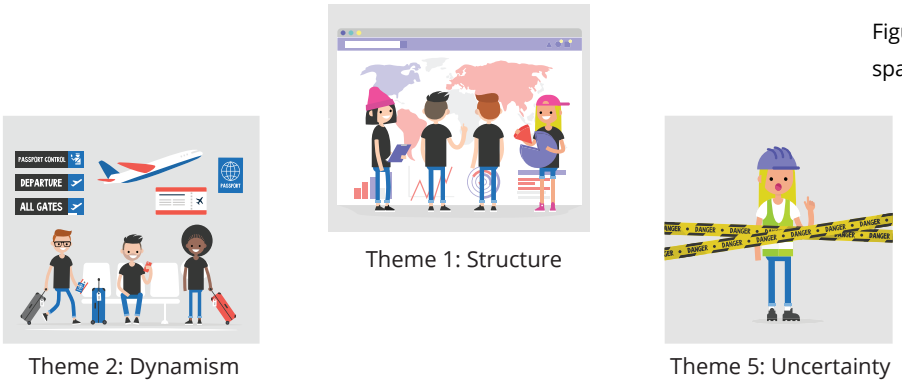
- GIS technology today is highly dynamic, with individual applications integrating many different components and systems together via a digital communications network. Chapter 7 explores the major concepts and technologies behind the architecture of GIS, including distributed, decentralized, streaming, and Web GIS.
- Where Chapter 7 concerns connecting computer-based components together, Chapter 8 covers the foundations of human user interaction with GIS. GIS can draw on principles common to design more broadly, in addition to the unique knowledge and centuries-old design traditions of cartography and map-making in interacting with spatial data.
- In Chapter 9, the focus turns to the combination of artificial intelligence and GIS. So-called "GeoAI" is a burgeoning topic today, but one that draws on long-established techniques and concepts in GIS and AI.
- Finally, Chapter 10 ends with a look at some of the most enduring problems in science that also present unique challenges to GIS, such as *uncertainty*, *privacy*, *fairness*, and *equity* in information systems.

We began this chapter with the question: what makes *GIS* special? We end with the broader question: what makes *spatial* special? Many books in the field of "geographic information science" (GI science), including this one, are linked by the central idea that spatial is indeed special. But while there are almost as many possible answers to the question of "what makes spatial special" as there books in the area, it is possible to identify five key themes that recur in unique combinations at the core of the field of GI science, summarized in Figure 1.25 (Duckham, 2015, and based on a published data analysis of the GI science literature).

1.5.1 Theme 1: Structure

A central theme running throughout this book—and indeed underpinning any book on spatial data—is that the *structure* of spatial information is differ-

ent to other types of data and requires special handling. We will encounter this theme in many different guises in this book (Figure 1.25, Theme 1): in Chapter 3 when we encounter Euclidean, topological, network, and metric spaces; in Chapter 5 when we examine the impacts of the jump from one to two spatial dimensions on computation and algorithms; and in Chapter 2 when we explore the effects of adding dependency between spatial dimensions on database indexes and spatial data retrieval. One pervasive example of the unique structure of spatial data encountered in Chapter 4 is *spatial autocorrelation*. But many other examples of the special structure of "spatial" have been proposed, such as statistical nonstationarity (Box 1.7 on the following page).



Theme 2: Dynamism

Theme 1: Structure

Theme 5: Uncertainty

Theme 3: Interfaces

Theme 4: Cognition

Figure 1.25: What makes spatial special?

1.5.2 *Theme 2: Dynamism*

Our geographical world is ever-changing. Nevertheless, static "snapshots" of our dynamic world are still adequate for many purposes, such as traditional cartographic maps for planning and navigation. Ignoring time, many of the foundational concepts and techniques covered in this book likewise concern storing and analyzing spatial "snapshots." However, we should not forget such snapshots are simplifications of convenience, and remain alert to the need for more sophisticated, integrated spatiotemporal concepts and techniques when it arises. Chapter 4 is the first point at which we introduce time into the equation, with inherently spatiotemporal phenomena such as movement (Figure 1.25, Theme 2). However, the importance of time, and the inherent

Box 1.7: Spatial nonstationarity

Spatial nonstationarity, sometimes also called *spatial heterogeneity*, concerns the statistical property of spatial data that no single "global" model can adequately capture its variability. In other words, there is no such thing as an "average" place. In contrast, many other data sets do exhibit stationarity, such as natural variation in human height, errors in GNSS positions, and economic processes and financial markets. Further, many other processes that are strictly speaking nonstationary, such as temperature time series, can often be transformed into stationary processes by removing trends or cyclic variations. Stationarity is important because an assumption of stationarity underpins many standard statistical tests and analyses. Because spatial data is nonstationary, such tests are often not well adapted to spatial data, and indeed the field of *geostatistics* has grown up around exactly this issue (see also Section 9.3.6). Goodchild (2004) discusses the importance of nonstationarity as a fundamental "law" governing spatial data, while any spatial analysis or geostatistics text will address the issue in more detail.

dynamism of any data about space, will recur at many points throughout the book and is a distinguishing feature of "spatial."

1.5.3 Theme 3: Interfaces

Spatial data and GIS technology pervade almost every aspect of our lives today, whether we notice it or not. From the biggest decisions of governments about how to respond to a global pandemic, to the smallest decisions about where to have lunch, spatial data is critical. The connection between data and decisions is provided by an increasingly wide variety of *interfaces*. To help people make the best possible decisions with the available data, those interfaces require careful and deliberate design. Happily, the design of interfaces to spatial data is one of the oldest and most well-understood areas of spatial expertise, embodied in maps and encompassed by cartography. The diversity of technologies that are used provides an interface to spatial data, from wristwatches to wheelchairs (Figure 1.25, Theme 3), means that modern approaches to spatial interface design must incorporate many more concepts and techniques than were required for hardcopy paper maps. Nevertheless, an understanding of the importance of interfaces, addressed in depth in Chapter 8, remains one of the central themes in "spatial" today, as it was from the earliest days of cartography.

1.5.4 Theme 4: Cognition

The contrast between humans and machines is nowhere more evident than when they must work in tandem to complete some task, for example, when flying a modern airliner. Aviation systems usually excel at reliably processing reams of data at lightning speeds; however, pilots must remain the masters of managing the unexpected, applying experience, adapting to context. Data is at its most valuable when systems help to combine these complementary strengths of humans and machines (Figure 1.25, Theme 4). Humans have evolved specialized capabilities for understanding and processing information about space. For example, like many animals, human brains possess *place cells* and *grid cells* that activate in specific ways when we move around a space,

helping us innately "know where we are." Our innate abilities for navigation and wayfinding, while not as astounding as salmon or shearwater, are nonetheless impressive. Ensuring effective collaboration between humans and GIS by acknowledging their complementary strengths and weaknesses is a running theme throughout the later topics in this book, including in connection with location-based services (Chapter 7), user interfaces (Chapter 8), and spatial reasoning and artificial intelligence (Chapter 9).

1.5.5 Theme 5: Uncertainty

Just as Theme 2 acknowledges that change is a constant in the world around us, Theme 5 acknowledges that uncertainty is endemic in spatial data about that world. We can never rely on spatial data with complete certainty. For example, all spatial coordinates must necessarily be limited in the level of detail given about a location, termed *precision*. Any observation of location is likewise subject to potential errors in measurement, no matter how small. An understanding of uncertainty is at the root of many areas of science and engineering, and geographic information science is no exception. However, uncertainty in "spatial" is concerned as much with techniques for managing and working with uncertainty—even using it to our advantage—as it is to reducing or expunging uncertainty, the focus in some other areas of engineering and the measurement sciences. Although uncertainty is endemic in all spatial data, the topic is addressed directly at the end of the book in Chapter 10, providing a framework for understanding the unique characteristics of uncertainty in spatial data.

Reflections

Introductory chapters to books are the most tricky to write, because they touch on many topics but dwell on few. Like this book, most other GIS textbooks, such as the excellent Longley, Goodchild, Maguire, & Rhind (2015), begin with some definitions but aim to give just a flavor of the variety of applications of GIS as well as of what makes spatial special. Other recommended introductory texts, including Heywood, Cornelius, & Carver (2011) and Burrough, McDonnell, & Lloyd (2015), similarly give their different perspectives on these topics.

Affectionately known as the "Big Book," Longley, Goodchild, Maguire, & Rhind (2005) is a more advanced two-volume collection of chapters by many of the pioneers of GIS, each of which delves directly into more detail on one topic. Despite its age, the first edition of this book is still relevant and at the time of writing some chapters are freely available online (Maguire, Goodchild, & Rhind, 1991).

Taking a slightly different tack from many other GIS textbooks, we chose to provide a more detailed summary of the computing technologies that underpin GIS, such as storage devices, networking, and the instruction cycle

in Section 1.4. This is partly because this book is, after all, "a computing perspective"; but also because we have found this provides a background and context that assists in understanding many features of information systems at the higher level of GIS (such as the relationship between power, range, and bandwidth in wireless networks, which helps when thinking about location-based services and wireless geosensor networks). Other recommended texts that provide more background on computer hardware, software, networks, and architecture include Tanenbaum (2013) and Null & Lobur (2018).

One topic that does not feature prominently in this introduction—but does in several other recommended texts such as Schuurman (2003), Chrisman (2006), Nyerges, Couclelis, & McMaster (2011), and Ballas, Clarke, Franklin, & Newing (2017)—is the importance of people and society in understanding GIS. That societal, human perspective is something we dip into throughout the book, in particular in Chapter 10. But the relative lack of coverage of this important topic is by no means a reflection of its importance, only of the focus and depth possible in a single-volume book.

One further difference between most other GIS texts and this book is the central focus on databases as the *engine* of GIS. This "database perspective" is at the heart of our "computing perspective." As a result, a solid grounding in databases is, we believe, a hallmark of the most capable GIS students and professionals, and it is the topic we turn to in the next chapter.