The core of GIScience
a process-based approach

Cover illustration:

Paul Klee (1879–1940), *Chosen Site* (1927)
Pen-drawing and water-colour on paper. Original size: 57.8×40.5 cm. Private collection, Munich
© Paul Klee, Chosen Site, 2001 c/o Beeldrecht Amstelveen

Cover page design: Wim Feringa

Contributors

This book is really the outcome of intensive collaboration within ITC. All those that have contributed in terms of the contents are mentioned at the individual chapters. At this place, we would like to acknowledge those people who have given an additional contribution. In the first place we like to mention the encouragement given by the previous rector at ITC, Martien Molenaar. The text is based on earlier versions of teaching materials to which many others have anonymously contributed. The LaTeX class used to produce this book was designed by Rolf de By. The figures have been compiled under the guidance of Wim Feringa, and author's editing was provided by Ian Cressie. We would like to acknowledge the contributions of them all. Finally, Teresa Brefeld, Ronnie Geerdink, Gerrit Huurneman, Fred Paats and many others are thanked for their practical support during the compilation process.

Valentyn Tolpekin Alfred Stein

Contents

List of Figures

List of Tables

Introduction

The concept of the Core

The core of GI Science: a systems-based approach is a textbook that aims to introduce students to the field of Geodata Processing and Earth observation. The book aims to represent the core of knowledge and techniques all students in the geoinformation sciences should possess. The book is suitable for use at all academic levels and should provide a basis for further study in more specialized directions. It reflects the core of scientific knowledge provided by ITC — covering all disciplines pursued within the faculty — this book is unique in being elementary, general and specific, all at once:

- *elementary* in the sense that students from different backgrounds, of varying levels, and from all parts of the world, should be able to understand it with no more than common sense knowledge of the field;

- *general* in the sense that it does not focus on a single domain of geoinformation science and/or Earth observation; and

- *specific*, as the fields covered by ITC are unique and specific in their focus.

For all these reasons, we refer to this book simply as *the Core*.

Why this book?

The core of GI Science: a systems-based approach originated as a textbook for introducing MSc students at ITC to concepts and techniques on Earth observation and geoinformation processing. The materials presented on its pages find their origin in the research and teaching activities of ITC, which spans a period of more than 50 years. ITC, the Faculty of Geoinformation Science and Earth observation of Twente University (Enschede, the Netherlands), operated until 2010 as an independent academic institution under the auspices of the Dutch Ministry of Education. The faculty's mission is, through the exchange of knowledge, to foster capacity building and the institutional development of professional and academic organizations, as well as individuals, specifically in countries that are economically and/or technologically less developed.

To facilitate this exchange of knowledge, ITC has over the years developed a curriculum to serve students coming from many different countries and representing a wealth of different academic backgrounds. Some of its students are already familiar with the most important basic developments in Earth observation and geodata processing and want to continue with more advanced studies. For them, this book establishes a common background of knowledge, thus providing a common language and understanding of the basic concepts. Other students have been active in the field for many

years, but lack a professional master's degree or MSc; they are eager to upgrade their knowledge and experience with a thorough understanding of the basic, state-of-the-art knowledge and insights from the field of Geoinformation science (GI Science) that consist of Earth observation and geodata processing. Still other students may possess a solid knowledge of computers but are not at all familiar with, for example, remote sensing and geodata processing. And there are even some students who have little knowledge, if any, of computers or their use. This book aims to serve them all.

The core of GI Science: a systems-based approach has three main objectives. The first is to help students learn how to generate information about Earth processes from remote sensing data. The second is to help students learn how to generate information about the Earth from data stored in Geographical Information Systems (GISs). And its third objective is to teach students how GI Science is being used within and across administrative agencies, as well as by policymakers and private citizens.

The subtitle *a systems based approach* refers to the development of systems thinking. The systems refer to the Earth or its components as natural systems, it refers to satellites as systems to collect data that are subsequently translated into information and it refers to Geographical Information Systems as systems for storing and analyzing layers of information. Therefore it is really a systems based approach, an approach that also requires users. All systems will be considered and addressed at the appropriate places in the book to follow.

The system Earth

Knowledge about GI Science (see also Figure 1) is presented in this book from the point of view of a system; the word "system" being a general term that we use to identify some of the fields of interest and their interactions. This will not only guide us in what to do, but also indicate what we should not do. This point of view may, thus, lead to new fields and to new combination of fields.

We call our system *system Earth*. It is our intention to understand this system from a geographic perspective. We will focus on those processes in the system that we consider important to understand, using available knowledge and tools to do so. Indeed, because we cannot understand it all, we will limit ourselves to only those processes that we can possibly observe, process and model through the methods and techniques that we have at our disposal.

In broad terms, we refer to the approach outlined in the Core as the GEOSS+ concept, referring to the acronym (GEOSS) for the Global Earth Observation System of Systems (see their website `http://www.earthobservations.org`). Similar to that of GEOSS, our concept is centred on a portal through which users of geoinformation express their desire to receive information, possibly collected from different sensors and stored in different layers of Geographical Information Systems. Such a portal essentially includes the user, the Earth processes that determine which information can be made available at a particular resolution in space and time, how skillful geoinformation processing can proceed, and, in particular, relates to the questions and expectations of a broad set of current and future users. The concept that we apply in the Core, however, needs to be broader than that of GEOSS, which focuses mainly on observing natural processes. Our concept includes social processes and the processing of information in models, as well as the essential role of the users and stakeholders.

The processes that form and shape the Earth are, therefore, important elements in this book. A basic distinction can be made between natural and human-induced processes. Most processes found in our *system Earth* have both a natural component and

a component induced by human activities. Sometimes it is difficult to separate these two. Thinking about deforestation, for example, we can easily see that there is a natural component (the forest followed by agriculture) and a social component (economic forces, urban spread). All processes have their own scales in space and time. They all have causes and driving forces, and in a number of cases they have actors that affect them, as well as those whom they affect. Furthermore, all processes can be approached at several levels of complexity. Flooding, for example, can be modelled simply by asking the question "Which area of land is flooded?", but a fuller analysis may also include the socio-economic aspects related to flooding, its consequences for biodiversity, its influence on risk management, and detailed spatial modelling. The Core covers several such processes, all embedded in GI Science. As a consequence, the line of logic followed is that of an Earth perspective: the way that these processes can be observed; the way that they can be modelled; and some of the relations with their users. Our basic approach is, first, to describe how the processes act; second, to explain how the data are generated; third, to elucidate how these data can be observed with sensors on board satellites and airplanes; fourth, to describe how the data can be stored in a geographical information system (GIS); fifth, to describe how models can constructed using these data; and sixth, finally, to show how to communicate information and results to users.

In the Core, we consider, in particular, those processes that can easily be observed by relatively simple remote-sensing methods and for which spatial implementation with deterministic, spatial models is straightforward. These models are usually generic enough; more complicated processes may require more specialized approaches beyond the scope of this textbook.

Given the concept of the Core, it is logical that the conceptual modelling of Earth processes with (simple) mathematical and logical steps has an important place. As, however, the term *model* is broad in its meaning, we must be specific about what we understand when we use the word in a geoinformation context. In this, we make a distinction between two types of models:

- *conceptual models*, e.g. models that by means of mathematical, logical or statistical methods approximate processes in the real world as closely as possible and that can be implemented within a GIS; and

- *observational models*, e.g. models that determine which data and images provide essential information about real-world processes.

Conceptual models are closely related to a specific point of view. A forest, for example, can be considered as vegetation covering some part of the Earth, but it can also be considered as a form of land use, for example for recreation or agro-forestry. Observational models require both individual observations and integration of data. It is useful to realize that not all data are by definition good observations for all processes. In the ITC concept of GI Science, the two models are closely related and, therefore, both are dealt with in this textbook.

Within our central concept of GI Science, there is also a clear role for users of models. We define users as those who are actively involved in using and understanding the *system Earth*, posing questions about it and looking for answers and solutions to problems; sometimes they are also the problem owners. The users we have in mind are primarily, but not exclusively, users from developing countries. They may be individuals or institutions, each acting at different scales and with a wide variety of interests. These users pose questions about the system and have a keen interest in the output of the modelling process.

Understanding

① System Earth

Temporal / spatial scales
Complexity

② Theory formulation

③ Model
Conceptual
Observational

④ answer

question

ODA/ITC related problems

User

Geoinformation

⑤ Observation

⑥ Data Integration

output

System Earth

sub system

☐ ITC specialization   ☐ non-ITC
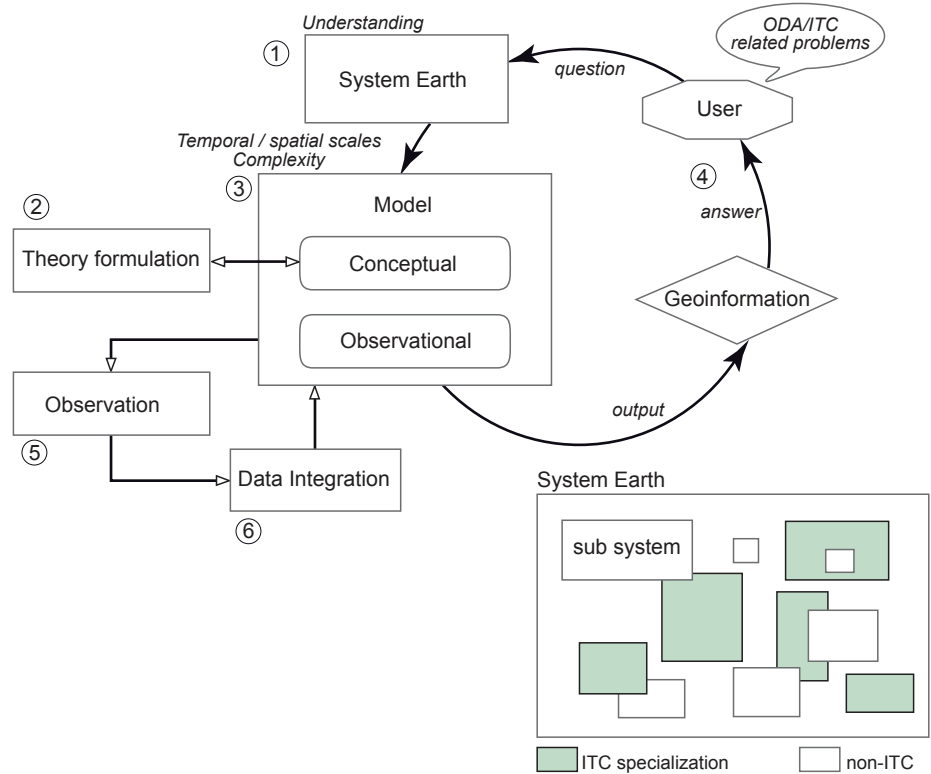
Figure 1
Diagram of the concept
followed in the Core

In the Core, we follow a number of steps in our systems-based approach to GI Science (for the following numbering, please refer to Figure 1). Starting with the system that is of central interest, i.e. *system Earth* (1), we then make the step from process to model. *System Earth* is modelled using *conceptual models* and *observational models* (2). The geoinformation processing chain (3)–(4) leads from the model to a GIS, from the GIS to *use of a GIS* and then back again to the model of *system Earth*. Here the user is using the information within a solid scientific model with a clear spatial component. We identify in this chain the step of theory formulation (3). Then, there is an important role to be played by the user (4), who develops a model, receives geoinformation from that model and poses questions about *system Earth*. *Observational models* include observations (5) such as field data, household surveys, time series of data, images, object-oriented observations, data on socio-economic processes, etc. The Earth observation chain (5)–(6) proceeds, therefore, from a *conceptual model* to an *observational model*, and from an *observational model* to *remote sensing*, thus feeding back into the model.

Users are involved in this entire process to a degree that is commensurate with their wish to understand or use Earth process—and hence their need for observations. At some stage (6), data from various sources and at different scales, collected at various moments in time, are integrated, for example as input for a conceptual model. All these steps are considered in the Core.

Of course, the central concept *system Earth* can be further subdivided into different subsystems. It is not so difficult, for example, to define the subsystem *Agriculture*, which includes many specialized systems such as *crop*, *soil*, *weed*, *groundwater*, *farming*, *economic management*, etc. Some of these specializations are within the ITC domain and are explored later in its educational program. Other specialities are just being developed and are only available outside ITC. In the Core, we aim to cover ITC's current specializations; "non-ITC" components are only included if they are considered necessary for a proper understanding of the processes that are inherent to the institute's mission. This notwithstanding, the Core still aims to be rich enough in content to enable readers to delve, if they wish, into non-ITC specialities.

Earth observation

To ensure progress in social well-being, sustainable economic development and environmental protection, we need to understand the natural and socio-economic processes related to our planet. Typically, studying a process involves observation, exploration, measurement and modelling. Such study is impossible without spatial and temporal measurements of the physical, chemical, biological and geometrical quantities that describe the process.

A major part of this textbook is concerned with the basic principles of *Earth observation*, which can be defined as the gathering of information about the physical, chemical, biological and geometrical properties of our planet; it helps us to assess its status and monitor changes in the natural and cultural environment. Earth observation gives us geodata—for mapping, monitoring, and also forecasting. Geodata acquisition can be taken as a starting point in our development cycle of observing–analysing–designing or planning–constructing or developing–observing, etc.

Earth observation by means of remote sensing (RS) has been defined in many different ways. A sufficient definition for this book is: *remote sensing* is the art, science and technology of observing an object, scene or phenomenon by instrument-based techniques. *Remote* because observation is done at a distance, without physical contact with the object of interest. We can either use detection and real-time display devices or recording devices of waves, which are emitted or reflected from an object or a scene. The waves can differ in nature: electromagnetic radiation (light), force fields, or acoustic waves are some examples. An example of a remote sensor is a conventional camera. Light reflected from an object passes through the lens and the light-sensitive film detects it. At the moment of exposure a latent image is recorded. Developing and fixing the film in the photo laboratory generates a definite record: the photograph. This image is then subject to interpretation. Today, most remote sensors are electronic devices. The data recorded by such sensors, e.g. a scanner detecting thermal emission (heat), is usually converted to images—for visual interpretation. Accordingly, we still refer to the record produced by an electronic sensor as an (remote sensing) image.

RS as defined above is applied in many fields, including architecture, archeology, medicine, industrial quality control, robotics and extraterrestrial mapping. Nevertheless, our domain of interest remains Earth observation, and in particular Earth observation from airborne or space-borne platforms. Earth observation not only relies on RS but also on sensors that allow us to make *in situ* observations. The principles of sensing in physical contact with an object are, however, beyond the scope of this book. In limiting the spatial interest of RS to objects that can be (and are relevant to being) located on the surface of the Earth—using a geodetically-defined coordinate system—we use the term *geospatial data acquisition* (GDA). The outcome of GDA is not simply an image obtained by converting sensor recordings, but is, rather, adequately

processed or interpreted data from the sensor. Processing/interpreting and validating require knowledge of the sensing process and yields data readily suited for analysis, e.g. in a GIS. Typical products derived from geospatial data include orthophoto maps, "satellite image maps", topographic maps, thematic maps (e.g. land use maps) and land use change statistics.

To place geospatial-data acquisition in a historical perspective, we can take surveying and mapping as starting points. About a century ago, photogrammetry evolved as a sub-discipline of surveying and mapping, offering an extension to traditional ground-based methods. Photogrammetry and other aerial surveying techniques could to a large extent replace observing directly in the field by measuring terrain features— indirectly—on images in the office. The next technological extension of GDA was remote sensing, which has enabled us to "see" phenomena that our eyes cannot. We can detect, for example, thermal emissions and display them in forms that allow us to analyse them with our eye–brain system.

historical perspective

Geodata processing

Once data have been collected, their processing becomes essential and it follows, there-fore, that a large part of this book is devoted to geodata processing. In its approach, geodata processing addresses the fundamental research principles upon which GISs are based [39]. Typically, GI scientists set out to solve problems with a geographical component. These problems may range from the relatively trivial (e.g. how to travel from A to B) to the complex (e.g. understanding climate change). The more complex the problem, the more likely the GI scientist will need to call in the help of other experts. In solving geo-related problems, one has all kind of questions to answer. These can range from those asking for specific information (How many kilometres is it from A to B?), to those that result in some insight (What does the road network look like?— to be able to decide on alternative routes), to those that result in knowledge (How does the time of day influence travel times?—combining road networks with commuting), to those that require reasoning (How can we solve traffic jams?—involving not only the GI Sciences, but also economics and politics, and thus requiring discussion and reasoning with others).
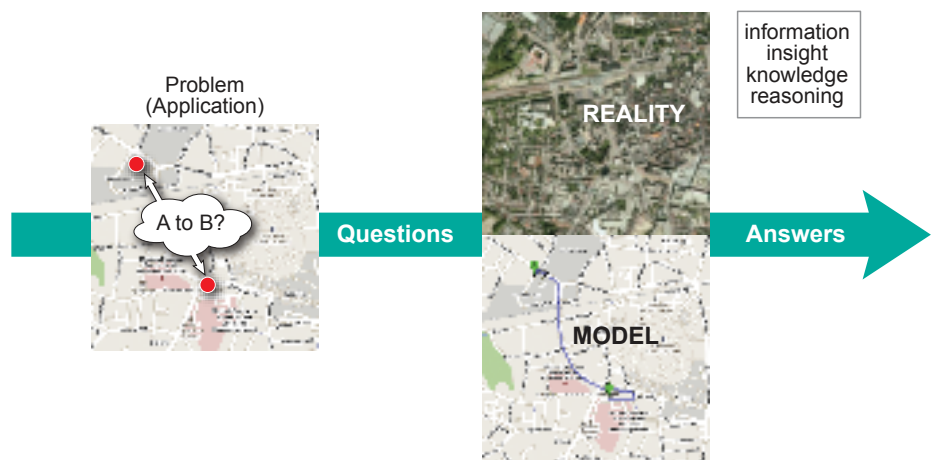
Figure 2
Modelling the real world to support the solution of geo-related problems. Source: [40]

To find answers to these sorts of problems, GI scientists have to simplify the real world—mostly because this real world is too complex. The complexity may be due

to the huge amounts of data available or perhaps a limited understanding of the real-world processes involved. For a simplified view of the world, GI scientists turn to models. A model is a simplified representation, description or simulation of reality. A map is a well-known example of a model. As a graphical representation, it is a selection (for instance, only roads) from and abstraction (for instance, symbolization of road classes) of reality. The objective of the GI scientist is to contribute to the solution of geo-related problems and prepare for or support decision-making.

For the process depicted in Figure 2 to be successful, some constraints have to be met: data are needed (see also Section 8.1); the process has to be organized (Section 8.4); a (Internet) working environment has to be defined (Section 8.5); functionality to run models and execute spatial analysis is required (Chapter 9); and results, often in the form of maps, have to be disseminated (Section 10.1).
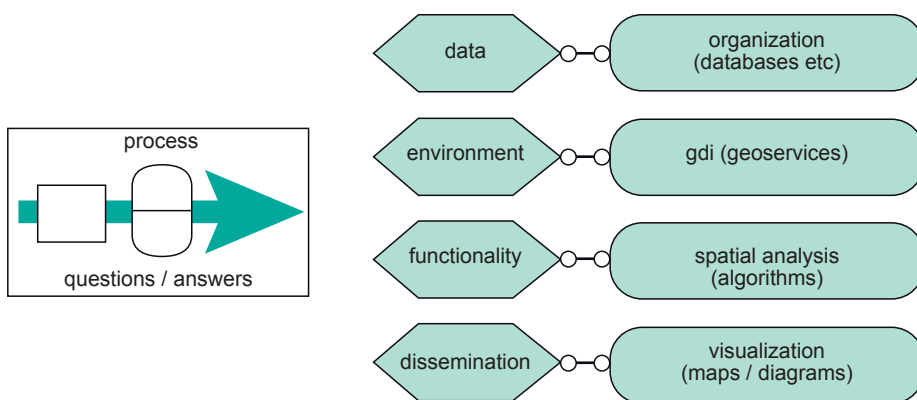
Figure 3
Modelling requirements and GISs.

These elements comprise the data and software part of a GIS. The user of a GIS, within a specific organization, is important as well. Organizational factors define the context and rules for capturing, processing and sharing geoinformation, as well as the role that a GIS plays in the organization as a whole. In Chapter 9 we focus on the architecture and functional components of GIS software.

To use data collected by Earth observation, the data have to be organized in such a way that one can access, query and integrate the data. Such data often have to be combined with other data—those obtained by surveying to measure parcels or prepare road works, for example. Also, different sensors may have been used, e.g. to obtain data for meteorological stations. Hence, integration of data is critical before adequate use of data can be made. And the type of user is important as well, putting their requirements on this integration. Geologists, for instance, need field work to understand complex Earth structures, and in the social sciences enquiries are a useful way of exploring options related, for instance, to the urban environment.

To answer questions, specific functions (tools) are needed to manipulate the data and so provide answers. Some of these functions will be described in Chapter 11, As such, they prepare data from different sources for common use (generalization). Elementary functions are those that identify and localize objects. In the "from A to B" example, particular roads could be identified by their name (e.g. the A1) or by their classification (e.g. motorway). These objects could also be measured, for instance the length of a particular road segment or the density of the road network. In general these functions could be grouped under the heading *measurement, retrieval and classification* functions. In the same example, the routing algorithm belongs to the category "connectivity or network functions".

Most functions combine different data layers, for instance a land use layer and a layer including a planned road to see what impact the new road would have on the environment. This can be done in vector or raster mode, depending on the nature of the available data. A combination of this sort belongs to the category overlay techniques. Neighbourhood functions are applied on a single data layer and their influence on the surroundings of the object of study are then considered. The impact of sound during the day along a road segment is such an example. Some functions that are considered to be interactive visual functions allow display and visual querying according to what is known as the Shneiderman mantra: overview first, zoom and filter next, then details-on-demand. Although these functions are apparently part of a graphic interface, they are in fact a combination of a visual and a computational approach. The algorithms take care of the "number crunching" and the human visual system detects patterns. Chapter 9 provides more insights into this.

During the last decade, the GI Science work environment has shifted from typical stand-alone applications to an Internet-based, networked environment. Data, and even certain functions, are offered as on-line geo-services that are part of a wider spatial data infrastructure (SDI). An SDI offers a set of institutional, technical and economical arrangements to enhance the availability (access and use) for correct up-to-date, fit-for-purpose and integrated geoinformation and geo-services. The services belonging to the infrastructure offer these arrangements in a timely fashion and at an affordable price to support decision-making processes. The motto is "collect once—use many times". Organizations within an SDI can be spread widely over several locations. With the development of the Internet, therefore, the functional components of GISs have gradually become available as web-based applications. Much of the functionality is provided by geo-webservices, i.e. software programs that act as an intermediate between geographic data(bases) and users on the Web. Such geo-webservices can vary from a simple map display service to one that involves complex spatial calculations. For their spatial data handling, these services commonly use standardized raster and vector representations, following the above-mentioned SDI standards. Section 8.5 escribes SDI in more detail.

Figure 4
The Geodata Infrastructure as working environment. Data can be available locally or accessible via Web Feature Services (WFS), Web Mapping services (WMS), via sensors (Sensor Web Enablement (WWE)). Web Processing Services (WPS) offer GIS functionality such as calculations or algorithms for specific operations.
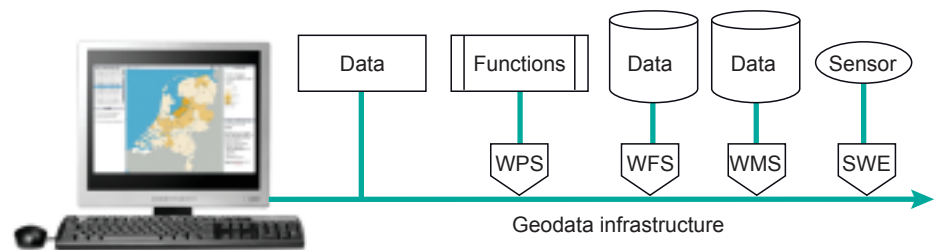
For reasons that include efficiency and legislation, many organizations work in a co-operative setting in which geographic information is obtained from, and provided to, partner organizations and the general public. The sharing of spatial data among the various GISs in those organizations is of key importance and aspects of data dissemination, security, copyright and pricing require special attention.

For GI scientists it is important that the data required for their research are available on demand, preferably via web services. Depending on the problem at hand, they can use the data in their own software environment, run spatial analyses or compile maps via webservices. Throughout this interactive process, GI scientists may need in-

termediate results, which may be in the form of sketches, maps, diagrams, tables, text, photographs and videos, but also notes and annotated documents collected during fieldwork or meta-data from obtained data sets. Their visualization may trigger ideas since they will show outliers or other spatial patterns that are likely to make geoscientists even more curious, perhaps leading to yet another new step in the spatial analysis process. Visual analysis always goes hand in hand with computational operations and is considered part of the modelling process. See also Section 10.1

The user

A special place in this book has been reserved for the user. Often we will consider the user explicitly, assigning him or her a well-defined position. The user poses both the questions that can be answered in a GI Science context (to decide upon the required observations and the analysis) and the interpretation and use of the outcomes. The users who might particularly benefit from GI Science are virtually infinite:

- In matters relating to food security it is essential to be able to forecast the overall agricultural production of a large area. An agriculturist therefore needs to know the size of fields per crop and have data on biomass production to estimate the total yield. Observing soil properties and monitoring degradation will improve forecasts.

- A user may be concerned about nature conservation. Nature conservation can only be well understood if natural vegetation patterns are clearly identifiable. A pattern of natural vegetation reflects the landscape, but also soil conditions and the presence of water. Under natural conditions, the processes that directly determine such patterns are, for example, the degree to which vegetation striving towards biodiversity, the effects of animal dynamics, as well as weather and climate processes. At a somewhat deeper level, subsoil and geology play an important role, influencing many of the other components. Apart from these, there are also the effects of human influence, as nowadays large areas of natural vegetation are maintained by humans. In total, we notice that a vegetation pattern is the result of a number of interacting processes.

- In health studies, an environment analyst may be worried about pollutants in waste disposal sites as these may affect the health of the local population. She or he therefore has to detect the disposal sites and determine their volume, as well as identify transport flows towards population centres.

- Nowadays many users are interested in the mapping of hazard zones. Politicians, urban planners and those working on infrastructure have their concerns about the safety of cities, roads and transport. Hazards are largely determined by geological phenomena such as earthquakes, volcanoes and landslides, which in turn are governed by geological and geomorphological processes. The effects are most dramatic when such natural disasters result in the loss of human life. Observing and modelling these processes may lead to a better predictability of the hazards and risks, and, thus, ultimately to a safer society. Roads and cities could be built where the risks are lowest, and rescue services could be sited near locations where the risks are highest. Clearly, social and natural processes are tightly interwoven.

- Mineralogists and mining engineers are interested in mineralogical patterns, generated by geological processes. Much of the interest has been based on the

seemingly perennial desire to discover gold, but new applications have evolved that focus on the search for other precious minerals (especially in non-vegetated areas). Although the processes that generate these mineralogical patterns have been going for eons, they may only be detected from satellites and survey aircraft. A solid investigation of the patterns may elucidate components of the processes that have helped to shape the Earth's crust.

- Environmentalists are increasingly concerned about the scarcity natural resources. Sea-level rise, global warming and the quality of the air are all components of our living environment that are quite clearly observable from sensors on board satellites and aircraft. By collecting repeated observations, combined with skilful modelling, these environmentalists can obtain detailed information about the rate and scale of changes (in space and in time) and which will enable them to separate the process from (random) noise. This may help to support political decision-making about measures to be taken.

- The International Panel for Climatic Change (IPCC), which is interested in the retreat of glaciers and shrinkage of the polar ice cap, is a particular example of environmental applications of GI Science. Most likely, these phenomena are caused by both human processes (CO_2 emissions) and natural processes (solar spots). Observing these events and understanding the processes involved can lead to a better understanding of how rapidly they are evolving—and how great our concern should be. This may be an important modelling activity in a GIS.

- Hydrological phenomena have been important for many years and are most likely to become even more important in the near future. In this broad context, a sub-Saharan farmer has a major concern in the form of variation in the supply of fresh surface water; dike-reeves and politicians are interested in the risks of flooding and the possible consequences for loss of land and life; environmental scientists are concerned about the conditions to which corals and phytoplankton are exposed in tropical coastal waters; the detection of sub-surface water in arid environments, which these days can play an important role in supporting growing populations; modern viewpoints have also assigned a large role to (sea-)water as a guarantor of supplies of food and drink.

- These days more than 50% of the world's population lives in an urban environment. Here, human-induced processes such as urban sprawl, and the traffic, housing and urban heat associated with it, are sources of major concern. Urban processes have been analysed and will become more and more important as populations increase, causing many problems in the major cities of the world: congestion of transport, deterioration of air quality and development of slums are only just a few aspects that we can observe at the Earth's surface, for which we need to acquire further understanding, and for which we can create useful models and development scenarios.

- Property boundaries can coincide with observable terrain features, which a land administrator needs to have surveyed. A land administrator should, therefore, have an up-to-date record of property boundaries. We focus on land administration that is either available and archived or that can be deduced from remote sensing images. For many years, a large number of GISs were based on some of the apparently simple linear concepts of roads and transport lines, the houses and blocks being considered as areal objects, and road and rail intersections as point objects. As the density of all these features increases, the need for GISs becomes more and more important, in particular in developing countries, where often land administration is still in its infancy.

- Spatial planners and industrial users of space are increasingly interested in complex linear systems, such as those of roads, pipelines and railway tracks. The networks they describe and their development in time can be difficult to understand but they seem to follow similar patterns. Again, natural aspects of system Earth largely determine where these linear systems *should* be situated, but often it is the societal aspects that determine where they *have* to be placed.

- A civil engineer who has to design a highway may look for criteria for finding an optimum alignment of the new road by balancing cut and fill volumes in road construction. Thus, he needs information about the shape of the ground surface. Calculation of actual transportation costs of material can be based on re-surveying terrain relief after construction. Alternatively, an urban planner may want to identify areas of informal settlement. The different types of houses and their configuration need to be determined. The municipality may furnish infrastructural improvements based on a development plan for the identified areas. The urban planner will have to monitor the impact of the provisions before proceeding further with any planning.

All the users and processes just mentioned have a range of similar characteristics. First, the processes all occur on the Earth's surface. Second, the results of those processes, i.e. the spatial patterns, can be observed with sensors on board satellites, aircraft and/or other vehicles. Third, the processes all differ in their spatial support and dynamics; i.e. all can be investigated at higher or lower levels of accuracy, and some can only be monitored by judicious selection of the time interval. Fourth, and finally, most of the processes involve both a physical component and a social component: no people, no problems!

Chapter 1

System Earth: some theory on the system

Emile Dopheide
Freek van der Meer
Richard Sliuzas
Anne van der Veen
Alexey Voinov

Introduction

"Everything is connected"—this statement is reiterated repeatedly as we witness some dramatic changes in the world that we live in. Our consumption of coal, oil, and other fossil fuels suddenly results in changes in our climate. Building bigger and more powerful refrigerators and putting air conditioners in our houses and cars, results in more skin cancer because of the depletion of the ozone layer. Converting mangroves into shrimp farms makes villages more vulnerable to storm surges and tsunamis. Building larger cities and paving more land unexpectedly results in less groundwater recharge and more severe flash floods and deteriorating water quality in our streams and lakes. The list of examples can go on and on.

In all these cases we are dealing with multiple factors that interact through something that Fritjof Capra [15] called a "web of life", an intricate collection of interacting nodes and links, actors and processes, causes and effects, stocks and flows that span across various spatial and temporal domains, with no recognition of disciplines, countries or even continents. In each case the extent, scope and strength of these influences and relationships can be different and has to be analysed and framed to understand them and to be able to foresee the possible effects and consequences.

That is where we start to speak about systems. How are elements are connected? What are the elements? How one influences another? Where are the causes and what are the effects? Bertalanffy [123] came up with the idea that there are similarities in how certain systems behave and that knowing some general principles we can be able to predict their future behaviour. But what is a system and what is system Earth? What tools do we have to analyse and govern system Earth? These are the main topics of this chapter.

While reading this chapter you do want to remember that by describing reality around us and by allocating words to the things and phenomena that we observe we immediately apply a certain frame, a lens, a point of view for our observations. Natural scientists prefer to think in terms of causes and effects, but for social scientists there are other possibilities, for example, meaning, storytelling or narratives related to the structure and agency of human behaviour. This goes back to an on-going discussion in science philosophy about the difference between natural and social sciences. There is no decisive answer but before we proceed with details on systems analysis, and before we discuss the big natural and human processes that are at stake on Earth, let us sketch some possible differences between natural and social sciences.

Science is about knowledge: finding explanations for the world around us and thinking about how we can come to predictions on the basis of our explanatory framework. Although in some literature there is a tendency to differentiate between science and social science (e.g. the Cambridge Dictionaries online defines science as (knowledge from) the systematic study of the structure and behaviour of the physical world, especially by watching, measuring and doing experiments, and the development of theories to describe the results of these activities), we will assume that "science" refers to something more general that includes all kinds of sciences and will specifically consider differences between natural and social sciences.

What is this difference, if any? One difference is in the concept of "laws". Most physical processes can be traced back to a few principles: conservation of momentum, mass and energy. Most scientific work in physics and applied fields like civil engineering build on those laws. But are there any laws in social sciences, based on which we can design experiments, run our models and present predictions? So far we can hardly think of any. Social science is about "meaningful action" and consequently, social phenomena that can be derived from the actions of purposive people either as individuals or groups and organizations.

However this distinction becomes less obvious if we recall that in biology, a natural science, there are also numerous interactions and processes that describe behaviour of organisms and species, and which would be difficult, perhaps impossible, to describe by the conservation laws from physics. Moreover, in quantum physics, for example, we also encounter phenomena that are yet to be explained in terms of first principles. Such similarities have inspired Capra [14] to parallel physics with Taoism ("The Tao of Physics"), demonstrating that in fact in complex systems there is much in common in how social and natural reasoning can be derived. While some, including Habermas, a famous philosopher, are of the opinion that systems analysis restricts social science, it is also conceivable that systems thinking may be useful in analysing social phenomena [12].

Nevertheless we should always keep in mind that systems approach is only one of the possible languages that can be used to describe the world around us. There may be many others, which would be based on other formalisms and other principles. Music, art, stories or narrative descriptions may be equally useful and powerful. Looking at reality through systems analysis we apply only one of the many possible lenses, or languages. By using it we are by no means diminishing the power and relevance of other languages. Indeed in this chapter we will examine to some extent the relationship between Governance and system Earth and in doing so will reveal some of the tensions that perhaps exist between systems analysis and the social sciences. But we also underline the need to at least attempt to reconcile such tensions, or at the very least be aware of their existence. But more of this later after we have described more fully what we mean by systems and some of the ways we have to model them.

1.1 Systems

Hall and Day [43] suggest that "any phenomenon, either structural or functional, having at least two separable components and some interaction between these components may be considered a system." Voinov [121] defines a system as "a combination of parts that interact and produce some new quality or function in their interaction".

Systems are everywhere. We just need to learn to identify and see them. This helps understand the world around us, and to better know what to expect in the future. For example we can see how application of fertilizers by individual farmers can entirely change the ecosystem of the river or lake downstream, or realize how changes in individual behaviour can impact the course of a whole economy, or how people respond to changes in government policy.

"The whole is more than the sum of parts." Bertalanffy

We should keep in mind, however, that systems are constructs of our brain. There is no single objective description of reality in terms of a system. The description chosen will depend upon the goals, the purposes of the study, the available information about the objects that we study, the level of detail, the scale and resolution that we choose.

There are three important characteristics of systems:

- systems are made of parts, or elements;

- these parts are in interaction;

- the interactions between elements result in new features of the whole.

All the three characteristics are essential for a system to be a system.

A collection of rods, wires, nuts, bolts and various pieces of steel of particular form that are put together in a special way make an internal combustion engine, which can now run your car. We can look at the engine as a system made of all these elements. Separately many of them are quite useless. Being put together and interacting these elements produce a new quality, or a new function displayed by the whole.

The choice of elements and their interactions in our presentation of a system will depend upon our viewpoint, upon what exactly we are trying to understand and what are we trying to achieve in the study. For example, we may describe a farm as a system where we have a crop growing in a field with inputs of water, fertilizers, etc. The elements would be the plants, the nutrient content of the soil, available soil moisture. These elements will be connected by flows of material between the plants and the soil, which will define how the plants grow. On the other hand, the same farm may be presented as an economic system, where we will be concerned with costs of plants, fertilizers, irrigation, labour, fuel, etc. We will end up with quite different systems that describe the same real world object.

Elements

A system may be viewed as a *whole* or as a combination of *elements*. Elements need to be properly identified to describe a system [122], however listing elements is not enough to present a system. How elements interact is crucial. An element is a building block of a system that can be also considered separately, with its own properties, features. If you cut a cake into pieces you would not call these pieces elements of a cake, because they have no particular features to distinguish them from one another in the cake—there may be any number of pieces, and as parts of a cake they are all the same. Besides, the pieces do not interact and do not offer any other properties except those delivered by the cake as a whole. The only difference is in size. Just a piece of a whole is not an element.

If you divide the cake differently, separating the crust, the filling and the topping, then you will get something quite different from the whole cake. It will make much more sense to call these parts elements of the whole. The taste and other properties of each of these elements may be quite different: the taste of the crust is not the same as the taste of the filling and not the same as the taste of the cake as a whole. Now there are ways to distinguish one element from another. The new taste or look that are created when these elements interact, are the important feature of the system.

Parts simply brought together do not necessarily make a system. Imagine ten people in a big room. They may be elements of a system in terms of being separable and looking differently and carrying some unique properties. However, just as a group they would hardly be called a system. However adding some rules of interaction, we can start a basketball game when the group of people becomes a system that exhibits some new additional properties. It is these properties that can draw thousands of people to watch the game. We get some additional *emergent properties* from the whole, which none of the elements possess.

emergent properties

In one case we can focus on a system as a whole and study the behaviour of elements in their interconnectivity within the system. This approach is called *holism*. Here it is the behaviour of the whole that is important and it is this behaviour that is studied. The behaviour of individual elements is of less interest and is viewed only as part of the overall system behaviour. On the contrary, *reductionism* is the theory that assumes that we can understand system behaviour by studying the elements and their interaction.

holism

reductionism

The population of Enschede, for example, in a holistic approach would be described by some general variables, like numbers, birth rate, mortality, education level. We may include more details also measuring religious preferences, consumption patterns, affluence, etc. But still we will be describing the population as a whole. Alternatively, we could present this same population as a collection of much smaller entities, say, age groups, or urban clusters, or even individuals, and then describe each of these entities separately. The properties of the Enschede population as a whole will then emerge from the joint performance of these separate entities.

These two approaches parallel analysis and synthesis. In the reductionist approach one can reduce the study of a complex system to analysis of smaller and presumably simpler components. While there are more components to consider, their complexity will be lower and they will be easier to experiment with and to analyse. However, this analysis may not be sufficient for understanding the whole system behaviour because of the emergent features that appear only at the whole system level. The holistic approach is essential to understand the full system operation. However if the behaviour of the elements is already well studied and understood it is easier to understand the whole system performance.

Interactions

Identifying and listing all elements (Figure 1.1a) does not describe a system in full. There may be many different ways in which elements may be connected or related to each other. The interactions, relationships between elements are essential to describe a system. The simplest is to acknowledge the existence of a relationship between certain elements, like this is done in a graph (Figure 1.1b). In this case a node presents an element and a link between any two nodes shows that these two elements are related. However in this diagram there is no evidence of the direction of the relationship: we do not distinguish between the element x influencing element y or vice versa. This relationship can be further specified by an oriented graph that shows the direction of the relationship between elements (Figure 1.1c). An element can be also connected to itself, to show that its behaviour depends on its state. We can further detail the

description by identifying whether element x has a positive or negative effect on element y.
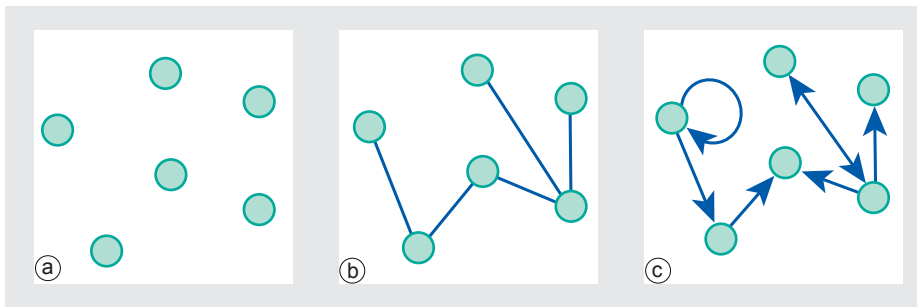
Relationships between elements can be described by two types of flows:

- flows of material, and

- flows of information.

Material flows connect elements between which there is an exchange of some substance. This can be some kind of material (water, food, cement, biomass, etc.), energy (light, heat, electricity, etc.), money, etc. It is something that can be measured and tracked. Also if an element is a donor of this substance the amount of substance in this element will decrease as a result of the exchange, while at the same time the amount of this substance will increase in the receptor element. There is always a mass, or energy conservation law in place. Nothing appears from nothing, and nothing can disappear to nowhere.

flows

The second type of exchange is with an information flow. In this case element A gets information from element B. Element B at the same time may have no information about element A. Even when element A gets information about B, B does not lose anything. Information can be about the state of an element, about the quantity that it contains, about its presence or absence, etc. Information flows can be used to describe rules and policies. Information flows can modify the rates of flow between elements, they can switch certain processes and interactions on and off. But the process through which policies, interventions and norms for action are established, and could for example define the values of such information flows, are themselves the result of social interaction between relevant stakeholders from public, private or civil society.

Note that when identifying the elements of a system and the interactions between them, we do not want to think that they are set in stone and are put in place once and for all. Systems can evolve. Elements can disappear, new elements and interactions can appear, the strength and even the presence of interactions can change in time. Policies and information variables may also change because we have a better understanding of a system works, but also perhaps because the stakeholders' consensus of what is important (i.e. their values) have changed.

Feedbacks

Interactions between elements may form loops. When A impacts B, while at the same time B impacts A, we say that there is a *feedback* in the system. Sometimes these loops are not that obvious and may involve several elements: $A \rightarrow B \rightarrow C \rightarrow D \rightarrow A$.

When describing flows in a system it is useful to identify stimulating or a damping effects. For example, consider a process of population growth. The larger the number of individuals in a population, the more potential births are occurring, the larger the number of individuals in a population, etc. This is an example of a *positive feedback*. There are numerous examples of systems with a positive feedback. When a student learns to read, the better she can read the more books she starts to read, the better she learns to read. Or another one: The more the weight of a person, the less fun it is for him to walk, the less he enjoys hiking or going somewhere, the less he burns calories, the more he gains weight. And so on.

positive feedback

On the contrary, a system with *negative feedback* tries to stabilize itself according to the rule: the larger something is—the smaller something becomes. For example, once again with populations: if there is a limited supply of food, and the population grows, there is less food for each individual. The more the population grows, the less food there is for each organism. At some point there is not enough food to support the population and some individuals die. Eventually growth shuts down completely, and the population equilibrates at a level that can be sustained by the supply of food.

negative feedback

Systems with positive feedback end up in uncontrolled exponential growth or collapse. Systems with negative feedback tend to stabilize at a certain level.

Structure

The elements of a system and their interactions define the system *structure*. The more elements we distinguish in a system and the more interactions between them we present the more complex is the system structure. Depending on the goal of our study we can present the system in a very general way with just a few elements and relations between them, or we may need to describe many detailed elements and interactions. One and the same system can be presented in many different ways. Just as with the temporal and spatial resolution, the choice of the structural resolution or the amount of details about the system that you include in your description depends upon the goals that you want to accomplish in your study.

Function

Whereas the elements are important to define the structure of the system, the analysis of a system as a whole is essential to figure out the *function* of a system. The function is the emerging property that makes a system to become *a system*. Putting together all the components of a birthday cake, including the candles on top, generates the new function, which is the taste and the spirit of celebration that the cake delivers. Separate elements have other functions, but only in this combination they create this new function of the system.

Defining system function can be tricky. The same combination of elements can result in different functions. Describing the interactions between elements in a particular system design is essential to define the function. Consider a birthday cake, where the right combination of cream, crust, nuts and fruits is essential to deliver good taste. The same cake can be used in a food fight to smash in the face of your opponent. The function is to offend and abuse your enemy. The taste in this case really does not matter, but what becomes important is the consistency, density and combination of solid and liquid elements.

The function is therefore determined by the "use" of the system, from the viewpoint of the analysis. What is the function of system Earth in this case? From the anthropocentric viewpoint, we can say that its function is to provide habitat and livelihood for human beings. From a more ecological viewpoint, we could say that its function is to sustain life in general. From an astronomical viewpoint its function might be to

provide the gravitational forces needed to keep the moon in orbit and to affect other planets and celestial bodies accordingly. We study systems for a purpose, and we describe systems in terms of those elements and interactions that are needed to best suit this purpose.

Delayed effects

Another feature often observed in complex systems are the so-called delayed effects. Systems and system elements do not always change immediately in reaction to external conditions or controls. With many elements and interactions involved, it may take considerable time for the signal to come through the system. Neither information nor material is transmitted instantaneously. As a result complex systems tend to behave counter intuitively, and they become hard to control. This is frequently observed in complex development related problems for which over simplified solutions are proposed. Another common phenomenon is that a change of policy might be claimed by its political advocates as having been very successful, when in fact the cause of success (desired change in a system's performance) might in fact be more rightly attributed to a myriad of other elements in the system, or perhaps even to a change in the environment in which the system is located. It is easy to overshoot the target if you judge only on the past behaviour of the system.

Hierarchy

Every system is part of a larger system, or a supra-system, while every element of a system may be also viewed as a system, or a sub-system, by itself (Figure 1.2). By gradually decomposing an object into smaller parts and then further decomposing those parts into smaller ones, and so on, we give rise to a *hierarchy*. A hierarchy is then composed of levels. The entries that belong to one level are assumed to be of similar complexity and to perform a somewhat similar function. *New emergent functions appear when we go from one level of a hierarchy to another*.
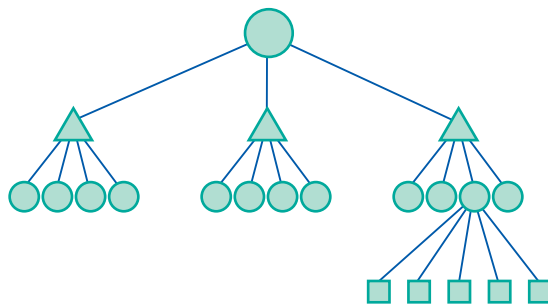
Figure 1.2
Hierarchies in systems. Whole systems may be presented as elements of another system and interact as parts of this supra-system. There are various hierarchical levels that can be identified to improve the descriptions of systems. Elements in the same hierarchical level are usually presented in the same level of detail in the space-time-structure dimensions.

For example, like other institutions, University of Twente has its own hierarchy, with the rector on top, followed by departments, staff, students, etc. Government typically also has a hierarchy from the bottom up: local, regional, national, and even some supra-national levels such as the European Union or the United Nations, though the status of supra-national bodies is not always recognized by national governments. Most private and civic organizations of any size will also have some form of hierarchy that is ideally established to facilitate the efficient and effective performance of its functions.

When analysing a system it is useful to identify where it can be placed in a hierarchy. The system is influenced by the higher levels and also by other systems at the same level. However, lower levels of those other systems are less important for this system.

They enter the higher levels in terms of their function; the individual elements may be negligible but their emergent properties is what matters. Fiebleman describes this in his theory of integrative levels as follows: "For an organism at any given level, its mechanism lies at the level below and its purpose at the level above" [32].

According to Saaty [101], "hierarchies are a fundamental tool of the human mind. They involve identifying the elements of a problem, grouping the elements into homogeneous sets, and arranging these sets in different levels." There may be a variety of hierarchies, the simplest are linear, such as:

universe→galaxy→constellation→solar system→planet→ · · · →molecule→ →atom→nucleus→proton

The more complex ones are networks of interacting elements, with multiple levels affecting each of the elements.

Again, it is important to remember that there are no real hierarchies in the world we study. Hierarchies are always an abstraction created by our brain and are driven by our study. It is just a useful way to look at the system, to understand it, to put it in the context of scale, of other components that affect the system. There is nothing objective about the hierarchies that we develop.

Systems thinking

More recently a whole new mindset and worldview based on the systems approach and systems analysis has emerged. It explores the interconnectedness between components and processes [15]. This worldview is referred to as "systems thinking". It looks at connections between elements, at new properties that emerge from these connections and feedbacks, and at the relationships between the whole and the part.

The roots of systems thinking go back to studies on system dynamics at MIT led by Jay Forrester. Back in 1956 he never mentioned systems thinking as a concept, but the models he was building certainly pioneered in holistic, integrative, cross-disciplinary analysis. With his background in electrical and computer engineering, Forrester has applied some of the same engineering principles to social, economic and environmental problems. There is a certain resemblance between electric circuits and systems diagrams that Forrester has introduced. The titles of his most famous books "Industrial Dynamics", "Urban Dynamics" and "World Dynamics" clearly show the types of applications that have been studied using this approach. The main idea was to focus on the system as a whole. Instead of traditional analytical methods, when in order to study we disintegrate, dig inside and study how parts work, now the focus was on studying how the whole worked, how the parts worked together, what where the functions, what were the drivers and feedbacks. But approaches have also had their critics. In the field of urban modelling for example, the work of Lee [66] was very critical of the value of early model builders' attempts to create models that could capture the intricacies of urban dynamics in a manner that was useful for urban policy and planning decision making, a criticism that for some is still valid today.

Later came the more sophisticated world models by Donegal and Dennis Meadows. Their book on "Limits to Growth" was published in paperback and became a national bestseller. System dynamics got a major boost when Barry Richmond at High Performance Systems introduced *Stella*, the first user-friendly icon-based modelling software.

Systems thinking is more than just system dynamics. For example the so-called Life Cycle Assessment (LCA) is certainly an example of Systems Thinking. The idea of LCA is that any economic production draws all sorts of resources from a wide variety of areas. If we want to assess the true cost of a certain product we need to take into

account all the various stages of its production, and estimate the costs and processes that are associated with the different other products that went into the production of this one. The resulting diagrams become very complex and there are elaborate databases and econometric models now available to make these calculations.

1.2 Models

A system is a *model*. We model all the time, even though we do not think about it. When we think we build mental models of the world around us. With words that we speak or write, we build models of what we think. It is sometimes hard to communicate because we are not always good at modelling our thoughts by words that we pronounce. The words are always a simplification of the thought. There may be certain aspects of the thought or feeling that are hard to express in words. Therefore the model fails and as a consequence we cannot understand each other.

The image of the world around us as we see it with our eyes is also a model. It is definitely simpler than the real world, however it represents some of its important features (at least we think so). A blind person builds a different model, based only on sound, smell and feeling. His model may have details and aspects different from the ones in the model based on vision, but both models represent reality in a simpler way, than it actually is.

The simplest definition of a model is that *a model is a simplification of reality*. As such it also needs to be an abstraction of reality, because in order to simplify we need to aggregate and describe the system in more abstract, general terms.

We tend to get very attached to our models. We think that they are the only right way to describe the real world. We easily tend to forget that we are dealing only with simplifications that are never perfect, and that everyone is creating their own simplifications in their particular, unique way for a particular purpose.

An example of a model we often deal with is a map. When your friend explains how to get to his house, he draws a scheme of roads and streets, in fact building a model for you to better understand the directions. His model will surely lack a lot of detail about the landscape that you may see on your way, but if it is a good model it will contain all the information you need to get to his house. If it is a bad model you lose your way and do not reach his house without additional help. For more sophisticated situations or to examine some of the "what if" questions that typify many development-related problems in which various alternative interventions might be investigated, for example, we can use Geographic Information Systems with spatial analysis and scenario building tools in which dynamics are becoming increasingly important.

Note that the models we build are defined by the purposes that they are to serve. If you only want your friend to get to your house, you will draw a very simple diagram, avoiding the description of various places of interest on her way. However, if you want her to take notice of a particular location, you might also show her a photograph, which is also a model. Its purpose is very different and so are the implementation, the scale, or the details.

Einstein is attributed the saying that "The best explanation is as simple as possible, but no simpler". The best model, indeed, should balance between realism and simplicity. The human senses seem to be extremely well tuned to the levels of complexity and resolution that are required to give us a model of the world that is adequate to our needs. Humans can rarely distinguish objects that are less than 1 mm in size, but then they hardly need to in their everyday life. Probably for the same reason more distant objects are modelled with less detail, than the close ones. If we could see all the de-

tails across, say, a 5 km distance, our brain would be overwhelmed by the amount of information it would need to process. The ability of the eye to focus on individual objects, while the surrounding picture gets somewhat blurred and loses details, probably serves the same purpose of simplifying the image, which the brain is currently studying. The model is made simple, but no simpler than we need. If our vision is less than 20/20 we suddenly realize that there are certain important features that we can no longer model. We rush to the doctor to bring our modelling capabilities back to certain standards.

As in space, in time we also register events only of appropriate duration. Slow motion escapes our resolution capacity. We cannot see how a tree grows, we cannot register the movement of sun and moon; we have to go back to the same observation point to see the change. On the other hand we do not operate too well at very high process rates as well. Even driving causes problems and quite often human brain cannot cope with the flow of information at the relatively high speed of movement.

Whenever we are interested in more detail in time or in space, we need to extend the modelling capabilities of our senses and brain with some additional devices: microscopes, telescopes, high speed cameras, long-term monitoring devices, etc. These are required for specific modelling goals, specific temporal and spatial scales.

The image created by our senses is static; it is a snapshot of the reality. It is only changed when the reality itself changes, and as we continue observing we get a series of snapshots that gives us the idea of the change. We cannot modify this model to make it change in time, unless we use our imagination to play "what if" games. These are the mental experiments that we can make. The models we create outside our brain, physical models, allow us to study certain features of the real life systems even without modifying their prototypes. For example, a model of an airplane is placed into a wind tunnel to evaluate the aerodynamic properties of the real airplane. We can study the behaviour of the airplane and its parts in extreme conditions; we can make them actually break, without risking the plane itself, which is many times more expensive than its model (see this web site for examples of wind tunnels and how they are used: `http://wte.larc.nasa.gov/`).

Physical models are very useful for "what if" analyses. They have been widely used in engineering, hydrology, architecture, etc. Mathematics offers another tool for modelling. Once we have derived an adequate mathematical relationship for a certain process, we can start analysing it in many different ways, predicting the behaviour of the real life object under varying conditions.

Consider Newton's second law:

$$F = ma$$

where F is a net force applied to a body of a mass m, and a is acceleration. The equation describes a simple model, which is a generalization of multiple experiments that show that for a body of mass m to accelerate at acceleration a it needs to be pushed (or pulled) with a force F that is a product of a and m.

This model is obviously a simplification of real movement, which may occur under varying force, mass, acceleration, where acceleration may choose direction, etc. However this simplification works well unless we get to speeds that are close to the speed of light, where this model fails and is no longer valid. By applying some general mathematical rules it may also result in additional findings, such as the relationship:

$$a = F/m,$$

which allows us to calculate the acceleration if we know the force and mass.

What is important in case of mathematical models is that some of the previously derived mathematical properties can be applied to your model in order to create new models, at little additional cost. In some cases, by studying the mathematical model you can derive properties of the real life system that have not been known previously. All models are wrong because they are always simpler than the reality. Therefore some features of the real life systems get misrepresented or ignored in the model. What can be the use of modelling then? Whenever you deal with something complex, you tend to study it step by step, looking at parts of the whole, ignoring some details to get the bigger picture. That is exactly what you do when building a model. Therefore models are essential to *understand* the world around us.

If we understand how something works, it becomes easier to *predict* its behaviour under changing conditions. If we have built a good model that takes into account the essential features of the real life object, its behaviour under stress will likely be similar to the behaviour of the prototype that we were modelling. We should always use caution when extrapolating the model behaviour to the performance of the prototype, because of numerous scaling issues that need be considered. Smaller, simpler models do not necessarily behave similar to the real-life objects. However, applying appropriate scaling factors, choosing the right materials and media, certain very useful results may be obtained.

When the object performance is understood and its behaviour is predicted, we get additional information to *control* the object. Models can be used to find the most sensitive components of the real life system. By modifying these components we can efficiently tune the system into the desired state or set it along the required trajectory.

In all cases, we need to compare the model to the prototype and refine the model all the time, because it is only the real-life system and its behaviour that can serve as a criterion for model adequacy. The model can represent only a certain part of the system that is studied. The art of building a useful model is mostly the choice of the right level of simplification that will match the goals of the study.

At this point we will give a general overview of the modelling process. This will be illustrated in numerous applications further on. Try not to get frustrated by the apparent complexity of the process and by some of the terminology that will be introduced with no due explanation. It is not as difficult as it may seem.

Building a model is an *iterative* process, which means that a number of steps need be taken over and over again. As seen in Figure 1.3, we start from setting the goal of the whole effort. What is it that we want to find out? Why do we want to do it? How will we use the results? Who will be involved in the modelling process? Whom do we communicate the results? What are the properties of the system that need to be considered to reach the goal?

We next start looking at the information that is available about the system. This can be either data gathered for the particular system in mind, or about similar systems studies elsewhere and at other times. Note that immediately we get into the iterative mode, since once we start looking at the information available we may very quickly realize that the goals we have set are unrealistic with the available data about the system. We need to either redefine the goal, or branch out into more data collection, monitoring, observation—undertakings that may shadow the modelling effort, being much more time and resource-consuming. After studying the available information and with the goal in mind we start identifying our system in its three main dimensions: *spatial*, *temporal* and *structural*.
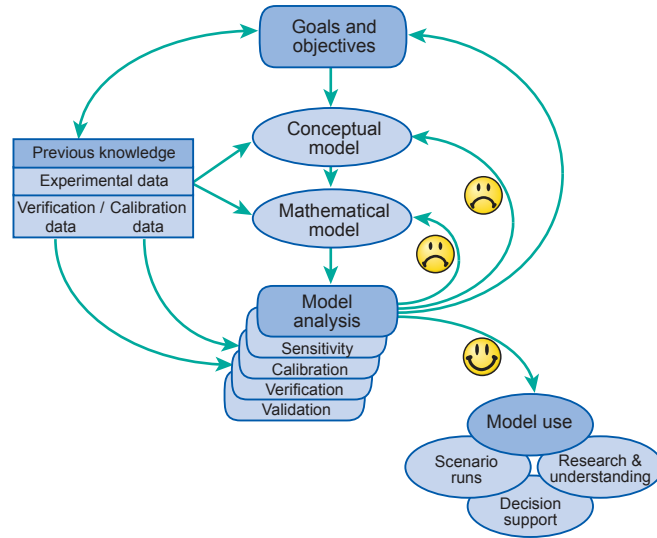
Figure 1.3
The modelling process. Note the iterative nature of the process: from almost any step we should be prepared to go back and start from the beginning, as we improve our understanding of the system while we model it.

Space   What is the specific size of the object that we need to analyse, in which system in the hierarchy our system is embedded in? How far spatially does that system extend? What will be the spatial resolution of the processes that we need to consider? How does the system evolve in space? Is it static, like a map, or dynamic, like the "game of life" (see `http://www.bitstorm.org/gameoflife/` or `http://www.mindspring.com/~alanh/life/index.html`)?

Time   What is the specific time of the system? Are we looking at it over years, days, or seconds? How fast are the processes? Which processes are so slow that they may be considered constant, which other processes are so fast that they may be considered at equilibrium? Do we need to see how the system evolves in time, like in a movie, or we just need a snapshot of the reality, like on a photo? If the system is evolving, how does it change from one state to another? Is it a continuous process or a discrete, instantaneous one? Is the next state of the system totally defined by its current one, or is it a stochastic process, where future states occur spontaneously with certain probability?

Structure   What are the elements and processes in our system? How much detail about them we need and can afford? Do we have enough information about all of them or some of them are entirely unknown? Which are the limiting ones, where are the gaps in our knowledge? What are the interactions between the elements?

We might already need to go back and forth from the goals to the data sets. If we realize that our knowledge is insufficient for the goal in mind, we either need to update the data sets to better comply with the goals, or we need to redefine the goals to make them more feasible at the existing level of knowledge.

By answering the basic questions about space, time and structure we describe the *conceptual model* of the system. A conceptual model may be a mental model, a sketch, or a flow diagram. Building the right conceptual model is half the way to success. In the conceptual model you should clearly identify the following components of the system:

Boundaries that distinguish the system from the outside world both in time and space. These are important to decide what material and information flows into

and out of the system, what processes are internal (endogenous) and which ones are external (exogenous). The outside world is something that you assume known and do not try to explore in your model. The outside world matters for your model only in terms of its effects upon the system that you study.

Variables characterize the elements in your system. These are the quantities that change in the system that you are analysing and reporting on as a result of the modelling exercise. Among variables we should distinguish:

- State variables, or output variables. These are the outputs from the model. These outputs are determined by inputs that go into the model, and by the internal organization of the model, the wiring.

- Intermediate or auxiliary variables, interactions, or flows. These are any quantities defined and computed in the model. Although they usually serve only for intermediate calculations, in some cases looking at them can help you understand what happens "under the hood" of your model. They are responsible for changing the state variables.

Parameters generally, are all quantities that are used to describe and run a model. They do not need to be constant but all their values need to be decided before the model runs. These quantities may be further classified into following categories.

- Boundary conditions. These describe the values along the spatial and temporal boundaries of a system. For a spatially homogeneous system we have only initial conditions, that describe the state of the variables at time $t = 0$, when we start the model, and the length of the model run. For spatially distributed systems, in addition, we may need to define the conditions along the boundary, as well as the geometry of the boundary itself.

- Constants or parameters in a narrow sense. These are the various coefficients and constants measured, guessed or found. We may want to distinguish between real constants, such as gravity, g, and, say, a population birth rate, μ. While both of them take on constant values in a particular model run, g will be always the same from one run to another, but μ may change quite substantially as we improve the model. Even if μ comes from observations, it will normally be measured with certain error, so the exact value will not be really known.

- Forcing functions. These are parameters that describe the effect of the outside world upon the system. These may change in time or space, but they do not respond to changes within the system. They are external to it, driven by processes in the higher hierarchical levels. Climatic conditions, rainfall, temperature certainly affect the growth of tomatoes in my garden, but the tomatoes hardly affect the temperature or the rainfall patterns. If we build a model of tomato growth, the temperature will be a forcing function.

- Control functions are also parameters, except that they are allowed to change to see how their change affects systems dynamics. It is like tuning the knob on a radio set. At every time the knob is dialed to a certain position, but you know that it may vary and will result in different performance of the system.

There may be a number of ways to determine model parameters, including the following [120].

1. Measurements *in situ* is probably the most reliable way since this can give the value of exactly what is needed for the model. However, such measurements can be very expensive and require a lot of work. They may also come with large margins of error. In some cases such measurements may not be possible at all, since some parameter represent aggregated values or extreme conditions that may not occur in reality (say, the maximum growth rate for a population under optimal conditions, which may not occur in reality).

2. Experiments in lab (*in vitro*) are usually performed when *in situ* experiments are impossible. Say we create optimal conditions for a plant to grow to find out its maximal growth rate. We can create such conditions artificially in a lab, but they may be impossible to find in natural ecosystems. Some would argue that experiments for social models are difficult to set up: imagine that we try out alternative tax systems to generate data on energy saving. Social scientists have to rely on surveys, analysis of historical events, and sometimes experiments in order to be able to understand and model human behaviour. Certain experiments with natural systems can also be simply impossible. For example, should geo-engineering be used as a solution for climate change?

3. Values from previous studies found from literature, web searches or personal communications. If data are available for similar systems it certainly makes sense to use them. However we should do it with caution since there are no two identical ecosystems or social systems, so most likely there will be some error in the parameters borrowed from another case study.

4. Calibration. This is when parameters are adjusted to make the model output fit the data as well as possible. See below for more.

5. Basic laws, say conservation principles and therefore mass and energy balances.

6. Allometric principles, stoichiometry, and other chemical, physical etc. properties. Basic and derived laws may help establish relationships between parameters, and therefore identify at least some of them based on the other ones already measured or estimated.

7. Common sense. This always helps. For example we know that population numbers cannot be negative. Setting this kind of boundaries on certain parameters may help with the model.

Note that in all cases there is a considerable level of uncertainty present in the values assigned to various model parameters. Further testing and tedious analysis of the model is the only way to decrease the error margin and deal with this uncertainty.

conceptual model

In most cases building a model starts with the design of a conceptual model. Creating a conceptual model is very much an artistic process, because there can hardly be any exact guidelines for that. This process very much resembles the process of perception that is individual for every person. There may be some recommendations and suggestions, but eventually everybody will be doing it in his/her own personal way. The same applies to modelling.

When a conceptual model is created it is often possible to analyse it with some tools borrowed from mathematics (sometimes it is not possible, especially when your concepts are qualitative). In order to apply mathematics you need to *formalize* the model, that is, find adequate mathematical terms to describe your concepts. Instead of concepts, words, images, you need to come up with equations and formulas. This is not

always possible and once again there is no one-to-one correspondence between a conceptual model and its mathematical formalization. One formalism can turn out to be better for a particular system or goal than another. There are only certain rules and recommendations, but no ultimate procedure known.

However, once a model is formalized, its further analysis becomes pretty much technical. You can first compare the behaviour of your mathematical object with the behaviour of the real system. You start solving the equations and generate trajectories for the variables. These are to be compared with the data available. There are always some parameters that you do not know exactly and that you can change a little to get a better fit of your model dynamics to the one observed. This is the so-called *calibration* process.

The relationship between your model and data is an interesting one. You should keep in mind that data are also a kind of a model (they are certainly a simplification of reality, and you certainly collect data for a certain process with a given resolution and accuracy). So in *calibration* you will be basically comparing two (or more) models: the *data model* that came from your monitoring and measurements and the *system model* that you have designed.

calibration

Usually it makes sense to first identify those parameters that have the largest effect on system dynamics. This is done by performing *sensitivity analysis* of the model. By incrementing all the parameters and checking out the model input we can identify to which ones the model is most sensitive. We should then focus our attention on these parameters when calibrating the model. Besides, if the model is already tested and found adequate, then model sensitivity may be translated into system sensitivity: we may conclude that the system is most sensitive to certain parameters and therefore processes that these parameters describe.

sensitivity analysis

If the calibration does not look good enough, you get a sad face and need to go back (reiterate) to some of the previous steps of your modelling process. Either you have got a wrong conceptual model, or you did not formalize it properly, or there is something wrong in the data, or the goals do not match the resources. Unfortunately once again you are plunged into the imprecise, "artistic" domain of model reevaluation and reformulation.

If the fit looks good enough you might want to do another test and check if the model behaves as well on a part of the data that was not used in the calibration process. You want to make sure that the model indeed represents the system and not the particular case that was described by the data that you used to tweak the parameters in your formalization. This is called the *validation* process. Once again if the fit does not match our expectation we get sad and need to go back to the conceptualization phase.

validation

However, if we are happy with the model's performance we can actually start using it. Already while building the model we have increased our knowledge about the system and our understanding of how the system operates. That is probably the major value of the whole modelling process. In addition to that we can start exploring some of the conditions that have not yet occurred to the real system and make estimates of its behaviour in these conditions. This is the "what if" kind of analysis, or the scenario analysis. These results may become important for making the right decisions. So a systems approach can be useful when we wish to introduce certain policy measures with the aim to change a system in a specific manner; a good model of a system can therefore help to identify efficient and effective interventions. We will come back to this later.

Complexity

When making all decisions about the model's structure, its spatial and temporal resolution, we should always keep in mind that the goal of any modelling exercise is to simplify the system, to seek the most important drivers and processes. However, there is always the temptation to make the model more complex. It seems that the more details you add to your model, the more precise is your description, the more accurate is your analysis and predictions. This is a risky path, because with more details you are also adding more uncertainty to the model. A complex model is also more difficult to test, to analyse. If the model becomes too complex to grasp and to study, its utility drops. There is little advantage in substituting one complex system that we do not understand by another complex system, which we also do not understand. Even if the model is simpler than the original system, it is quite useless if it is still too complex to shed new light and to add to the understanding of the system. Even if you can perform experiments on this model that you might not be able to do in the real world, is there much value in that if you cannot explain your results, figure out the causes and have any trust in what you are producing?

"It is better to be vaguely right than exactly wrong"
Read, [97].

1.3   Some simple models

There are many examples of models that are relevant for spatial systems analysis some of which will be described here. More attention is given to modelling and GI Science in Chapter 7. There are also various ways we can draw classifications of models (see [121] for some examples). One quite common distinction is between process-based and empirical or statistical models. In empirical models the output is connected to the input by a mathematical formula, and its form is not important as long as the input signals are translated into the output ones properly. That is, as they are observed. These models are also called *black-box models*, because what matters is only what is in the input and the output; what happens inside, in between, is of no physical meaning. For example, if a process is approximated by a polynomial, there is no meaning in the coefficients, they cannot be measured in reality, they can be only calibrated. In process-based models individual processes are analysed and reproduced in the model. Certainly, we will never be able to go into all the details and describe all the processes in all their complexity (it would not be a model then). Therefore, still, a process-based model will also assume a certain level of empiricism. It may be considered as built out of numerous black boxes. The individual processes are still presented as closed devices or empirical formulas, however their interplay and feedbacks between them are taken into account and analysed.

empirical models

Another common dichotomy is between deterministic and stochastic models. In a deterministic model the state of the system at the next time step is entirely defined by the state of the system at the current time step and the transfer functions used. In a stochastic model there may be several future states corresponding to the same current state. Each of these future states may occur with a certain probability. Some of these models will be considered later in this book.

deterministic and stochastic models

Conceptual Model

As mentioned above, in most cases any modelling process starts with a conceptual model. A conceptual model is a qualitative description of the system. A good conceptual model is half the modelling effort. To create a conceptual model we need to study the system and collect as much information as possible about the system itself, and about similar systems studied elsewhere. When creating a conceptual model, we start with the goal of the study and then try to explain the system that we have in terms that

would match the goal. In designing the conceptual model, we decide what temporal, spatial and structural resolutions and ranges are needed for our study to reach the goal. Reciprocally, the conceptual model eventually becomes important to refine the goal of model development. In many cases the goal of the study is quite vague, and it is only after the conceptual model is created and the available data sets are evaluated that the goals of modelling can become clear. As we see in Figure 1.3, modelling is an essentially *iterative* process. We cannot prescribe a sequence of steps that takes us to the goal. It is an adaptive process when the target is repeatedly adjusted and moved as we go, depending both on our modelling progress and on the external conditions that may be changing the scope of the study. It is like shooting at a moving target. We cannot make the target stop to take a good aim and then start the process. We need to learn to readjust, to refine our model as we go. Building a good conceptual model is an important step on this path.

In Figure 1.4 you can find an example of a simple conceptual model that can be used to explore and communicate our ideas about how a coupled human–natural system might operate. We choose certain elements of the system (Population, Economic development, Environmental Degradation, Investment) and add interactions between them. We can then start thinking about which processes are important in describing the changes in the elements, and how information and material are flowing in the system.

Fuzzy Cognitive Maps

A Fussy Cognitive Map (FCM) can be seen as the next step in formalizing a conceptual model. It consists of nodes (or concepts), which are the elements, or variables, with connections (or edges) between them that represent the causal relationships between the concepts. Each connection gets a weight (between 1 and 0) according to the strength of the causal relationship between the concepts in nodes (Figure 1.5). In this example, the role of human agents and their actions is implied in many of the nodes and their relations.

Implicitly, the model requires decisions to be taken about which elements or activities to regulate through enforcement, what norms or standards to apply, and what sanctions to impose. Such matters are resolved through governance processes or systems. A relationship can be either positive (when growth in one concepts stimulates growth in the other one) or negative (when growth in one concepts inhibits growth in the other one). This graphical form can be represented in a mathematical form of a matrix that lists all the connections and a state vector made of the current weights of the concepts in the system. The next state of the system is then calculated by multiplying the vector by matrix. If iterated, the system may equilibrate, presenting a new state. Comparing the new and initial weights of variables can tell whether the system increases or decreases their importance through the connections. FCM focuses on feedbacks within a system. They can help frame the discussions with stakeholders, whose expert opinions can be used to put together the initial maps.

FCMapper is a free software package to process FCMs and calculate important metrics for them. It can be downloaded for free from `http://www.fcmappers.net/`.

Maps

As discussed above, maps or images are also models in themselves. They are especially appropriate to describe the spatial arrangement and organization of the system. Maps can provide us with important information about the spatial relationships between various elements of the system. Maps come in different resolutions and granularity. As always with models, they are largely defined by their purpose, the goal of

the model. Depending on the goal, maps of the same spatial domain may look very different. Figure 1.6 presents two maps of the area around ITC (Enschede). One is a street map that one would use to describe the directions to ITC. Here we are not interested in various landforms and landscape features. What matters are the streets to choose the best route. The other model is an aerial photo, which does not have streets names on it, but has many landscape features clearly visible and can be used to calculate the areas of green space, calculate the impervious surfaces, identify the location of particular houses and so on.

System dynamics

The main assumption in such models is that systems can be represented as a collection of stocks connected by flows, so material or energy accumulates in stocks and moves between them through flows. Stella was one of the first software packages that captured worldwide recognition due to a very nice graphic user-friendly interface (GUI) and a fairly wise marketing program that particularly targeted students and university professors (Figure 1.7). A number of other software packages followed that are superior to Stella in many aspects. Other packages in this category are: Vensim, Powersim, Madonna, Simile and others. An expanded version of these are extendable tools, such as Extend, GoldSim, Simulink, and others. This software has many more icons than the stocks, flows and parameter operations of the basic Systems Dynamics tools. Whole sub-models or solvers for mathematical equations, such as partial differential equations, may be embedded into specially designed icons that later on become part of the toolbox for future applications. Clearly a major advantage is that modelling systems can be extended to include almost any process. However there are always limitations. A major one is that, systems dynamics tools in most cases are not well suited for spatially explicit formalization. Simile is probably the only one of those
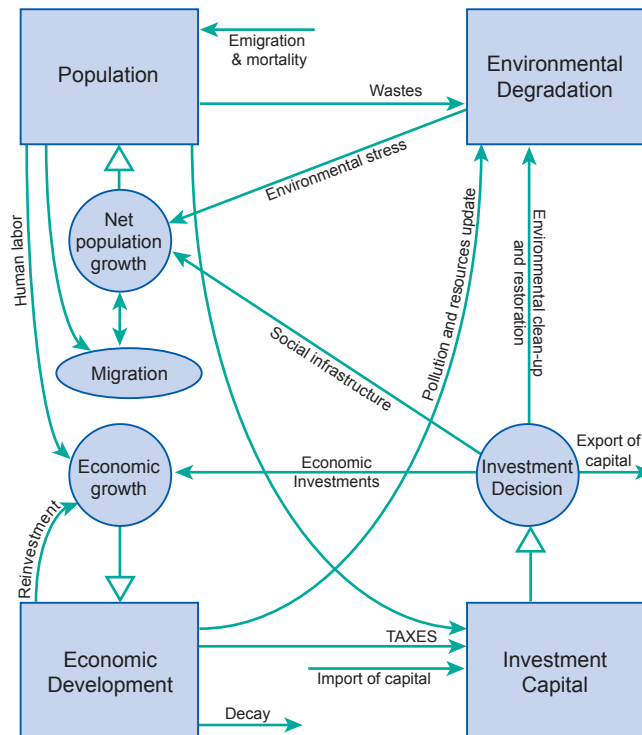
Figure 1.4
Conceptual model of a socio-economic and ecological system designed to analyse sustainable development.
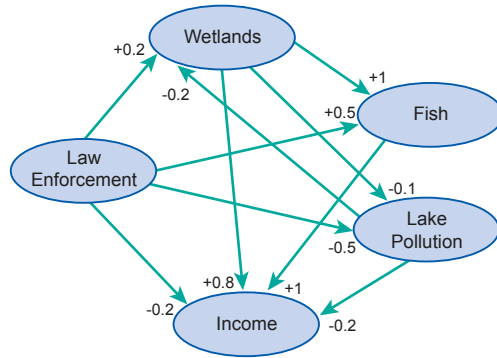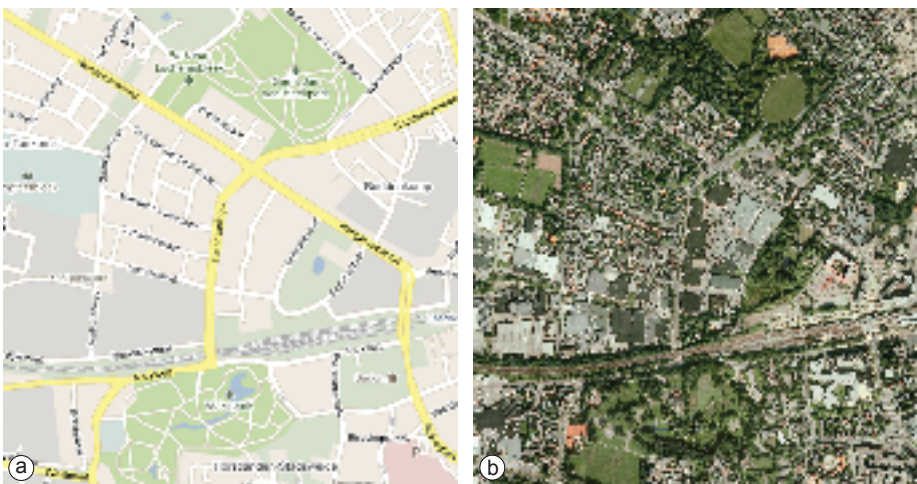
mentioned above that can handle maps and spatial transport well. Alternatively, Stella local models can be embedded in more sophisticated software tools such as SME [34]. More recently a translator was built to convert Stella equations into R script, making them available for further analysis and runs within the powerful open source R package [78]. One generic problem is that the more the power built into these systems, the more difficult they are to learn and to use.
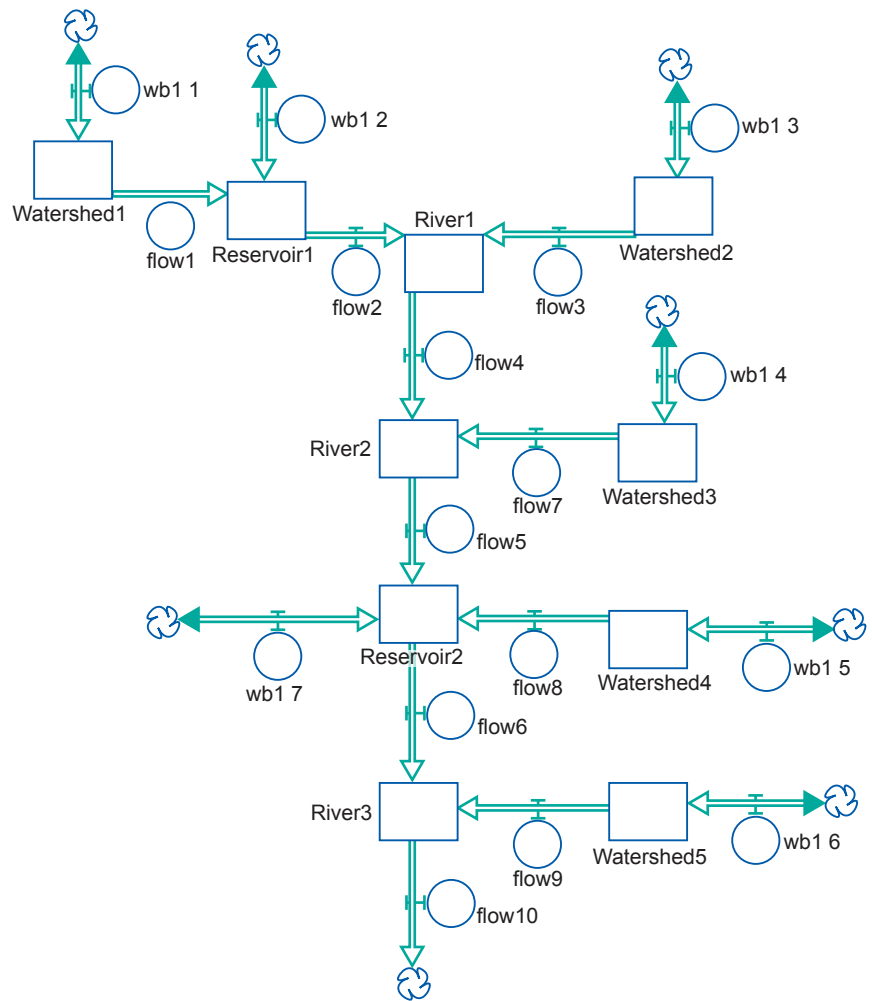
Figure 1.7
A Stella model of a cascade
of reservoirs and their feeding
watersheds.

Agent-based models

To model complex phenomena that involve human or institutional behaviour it is helpful to represent them as multi-agent systems (MASs) and use an Agent-Based Modeling (ABM) approach. MASs describe the observed world in terms of actors (agents) that are characterized by certain rules (behaviour) that depend on the state of the environment, the state of the agent and its spatial location. Each agent is represented as an independent computerized entity capable of acting locally in response to stimuli or to communicate with other agents. An example of an output from an agent-based model is presented in Figure 1.8, which shows how various agents, representing land owners, are distributed around the town centre on a particular coast, where the coastal amenities serve as an attraction and increase the price of the property, while at the same time they come with higher flooding risks, which make those areas inappropriate for agents with higher levels of risk perception. The applications are generally developed with an object-oriented language. Swarm [74] was one of the first software packages designed for ABM. Now tools like Repast, NetLogo, Mason or Cormas are more often used. These tools, which include spatial representation, simulation utilities for Monte-Carlo type methods, and links to other software (GIS, databases), are useful for the implementation of different social or ecological systems. Algorithms or structures are provided to implement the link between agents and their environment and routines are provided to organize societies of agents. During a participatory process stakeholders help to identify the actors involved in the decision making process, the interactions among them and with their environment. For the last decade, many scholars have considered a new implementation model for agent-based modelling. They implement them through Role-Playing games [11, 6]. The combination of computerized ABM and Role-Playing games can also be invoked.
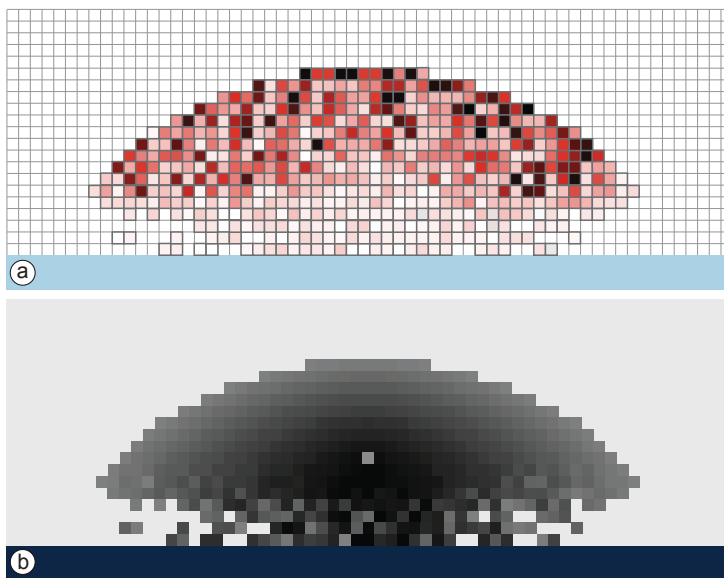
Figure 1.8
Results of an agent-based model for a land market in a coastal town with heterogeneous agents. (a) rent gradients, (b) risk perception of flooding. [33].

1.4 System Earth and governance

1.4.1 System Earth

The ultimate challenge in applying systems theory might be to design a working model of system Earth at various levels of complexity. Such a model would increase our understanding of the planet as an integrated system of components and processes. The modelling of system Earth is therefore a learning process, one which may expand and deepen our understanding and, one would hope, our capacity to manage environmental change and achieve sustainability.

Earth as a system may be described in terms of four principal components or subsystems: atmosphere (air), lithosphere (land), hydrosphere (water), and biosphere (life). Some breakdowns also include the cryosphere, specifically referring to those parts of the Earth that are dominated by ice masses, and the anthroposphere, which distinguishes the specific role of humankind as a part of the Earth's biosphere. While useful for some descriptions, these subsystems operate in close connection and are intimately interdependent.

All Earth's subsystems are driven by energy, the main sources of which are the Sun (heat, chemical energy stored in various chemical elements, etc.) and the Earth's core (geothermal energy and radioactive energy). Since solar radiation reaches the Earth at an angle, the distribution of radiation and, thus, heat over Earth's surface is uneven. This creates Earth's three major climate zones—tropical, temperate, and polar—which in turn drives life forms in these different regions. In addition, the uneven heating drives the Earth's weather. Heat absorbed by the oceans and carried by its currents is constantly being released into the atmosphere. This heat and moisture drive atmospheric circulation and set weather patterns in motion. These weather patterns influence, in turn, vegetation, as well as erosion and sediment transport. Radioactivity in the Earth's core produces a heat source that also is unevenly distributed throughout the Earth's mantle and crust. This heat source drives plate tectonic processes, which in turn are responsible for the topographic architecture of our planet and are associated with volcanism, earthquakes and other disaster-related geological processes.

climate change

Newspapers today are full of reports about melting ice caps, rising sea levels and higher than average temperatures, all of which can be connected to anthropogenic forces (human activity) that, among other things, produce carbon dioxide, which contributes to global warming. Excluding anthropogenic forces, the Earth in its geological past has also experienced warmer and colder periods, as well as periods of more and less CO_2 in the atmosphere. The Earth has a relatively warm surface temperature because gases in the atmosphere absorb heat, thus acting as a blanket over our planet. Sunlight can pass through the atmosphere while its heat is prevented from escaping, like in a greenhouse. Although there are several greenhouse gases (GHGs), carbon dioxide is the one climate experts are most concerned about because the unprecedented increase in human-related CO_2 emissions since the industrial revolution has now been linked to the current trend of global warming. Natural CO_2 emissions and sinks are largely in balance; it is human greenhouse-gas emissions that have disturbed this equilibrium and resulted in a rapid accumulation of CO_2 and other GHGs in the atmosphere. This has created conditions that can result in climate change. Multi-level governance processes are critical to an effective response to this situation but are so far largely failing to deliver effective responses at scale that will have the necessary impacts.

In Section 1.5 we will briefly examine some of the major processes that control system Earth, combining them in cycles that describe the dynamics of carbon, nitrogen and water. In addition to these natural processes, we will also examine the process

of urbanization, which might be seen as example of an anthropogenic process. It will become obvious, though, that in all of these processes there are important interactions between natural and anthropogenic components and, moreover, that governance processes are fundamental to the management of these processes and that an integrated, multidisciplinary perspective of these processes is therefore essential for this reason, we first need to look more closely at the issue of governance.

1.4.2 Governance

The natural components of system Earth are complex. This complexity is, however, further increased when we include the presence and actions of humans. Whether as individuals with "free will", or organized in various kinds of groups with a degree of collective action, humans are hard to understand, and it is difficult to forecast their actions. For a long time, humans were not such an important factor for the Earth as a whole. However, at present, humans have become a "geological power" [119]. We can already see that the extent of human activities on this planet can rival geological processes. Some have even suggested that this period of geological time should be termed the "anthropogene". Humans are already changing the climate of the Earth; they build dams that can be a factor in earthquake patterns; they are causing a loss of species that is going faster than at any time in the last 65 million years, when the Earth was hit by an enormous asteroid that wiped out thousands of animals and plants, including the dinosaurs. It is of paramount importance to understand the behaviour of humans on the Earth if we are to understand and predict the evolution of system Earth.

Individuals make decisions and behave on their own. They also have the power of *collectivity*, when they act together to reach certain goals that are beyond the power of individual action. Examples of collective actions are those that result in the production of public goods: for example, public infrastructure, police forces, dikes, environmental quality standards, and spatial planning. It is usual to aim to meet two criteria in the production of public goods:

- to do the right things, and

- to do things right.

The first criterion relates to the legitimacy of government, whereas the second refers to its efficiency and effectiveness. Governing society is not merely a principle of command and control, but is also a process of seeking a balance between powerful groups, on the one hand, and giving protection to the not-so-powerful on the other. To that end, we use the term governance.

Formally, we might define governance as all those interactive arrangements in which:

- public and private actors participate, with the aim of

- solving societal problems or creating societal opportunities, attending to the institutions in which these governance activities take place, and stimulating *normative debates* on the principles underlying all governance activities.

In highlighting the issue of normative debates, it needs to be stressed that in all discussions on the role of government choices have to be made, not only by government, not only by experts, but also by individuals and groups in society that have a stake in the outcome. Normative debates are the dialogues through which the essential values and aspirations of stakeholders are discussed and agreed upon. The outcomes form a basis

for plans, decisions and, ultimately, collective action. But there are no simple problems and simple solutions. Of course, building a bridge, or designing an irrigation scheme has to follow a certain logic. However, most societal issues are "wicked" problems in which different stakeholders or different groups have different views, values, stakes and power. For individual stakeholders there is also often much to be gained, or lost, as a result of the decision-making process. The distributive nature of public-policy choices is therefore significant and frequently controversial. All this only adds complexity to the system we are studying.

How do we then handle the role of government, or, in more general terms, governance, in a systems approach? As we have already noted in Figures 1.4 and 1.5, changes in (legal or economic) instruments designed by government may influence individual and group behaviour. Groups in society may adapt to a more powerful position (e.g. Tunisia, Libya); people may protest and raise their voice against government (as they currently are in Greece). This may change the structure and scope of our models and/or their parameters.

1.4.3   The policy cycle

A simple way of introducing governance issues into a systems analysis framework is by incorporating what is known as the policy cycle, which specifies the stages of policy design and policy evaluation. As a first step, to recognize the importance of governments, it is instructive to briefly discuss all the various stages.

Government, business and industry use policies to guide decision-making in the real world. In government, public policy defines and structures a rational basis for government action. Public policies can be classed as distributive when they extend goods and services to members of an organization or society. When limiting the freedom (in the broadest sense) of individuals or organizations, policies can be classed as regulatory or mandatory, while they can be classed as constitutional if certain entities can derive power from the policies.

The scope of public policy determines the people or target groups, organizations or nations, and the elements within the natural environment, that the policy affects. Policies often have unintended consequences, which should be avoided. Consequently, a learning stage provides for monitoring and evaluation of a policy instrument after careful formulation and implementation, which may result in re-formulation of the policy if the policy measure does not appear to be effective, or if undesirable side-effects are too burdensome. Note that we made the comment earlier in this chapter that one of the differences between the natural sciences and the social sciences is the role of experiment. It is extremely difficult to set up experiments aimed at testing the social dimensions of certain incentives. Therefore, it is worthwhile to look at policy design and policy implementation as a cyclical process, where we slowly try to understand how individuals and groups will react. Such a policy cycle (Figure 1.9) is a heuristic tool from political science that can be used for analysing the development of a policy following certain steps, or stages: policy formulation (determination and description); policy realization (implementation and dissemination); and policy learning (monitoring and evaluation).

As an example, in Earth and life sciences, we have a mandate to be concerned about our environment, a notion that is laid down in environmental policy- making. Space-borne remote sensing and environmental policy came of age at the same time. The launch of the first Landsat satellite in the early 1970s coincided with the development of the first environmental policies, and the environmental protection agencies that were subsequently established also created the initial demand for environmental Earth Observation (EO) products. In this context, the following subsection describes
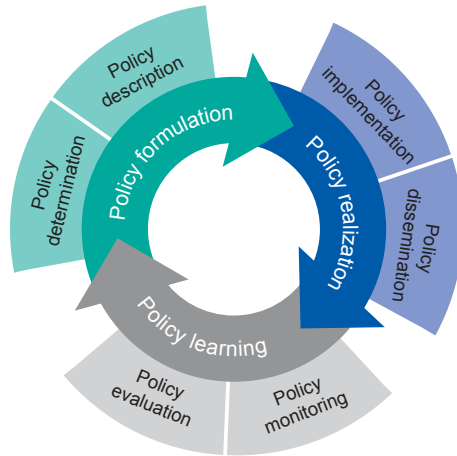
some examples of the potential of EO in three particular stages of the environmental policy cycle: problem determination, policy implementation; and policy monitoring. Some examples of EO applications are also summarized in Table 1.1.

	Environmental policy questions per stage of the policy cycle	Examples of potential/actual EO contribution to the policy cycle stage
Problem identification	Is P emitting E? Does E cause harm? Can V perceive E?	Size and depth of ozone hole Deforestation Polar sea ice cover reduced
Causal specification	*Does P's E affect V?* *Does A cause B?*	*Dredging causing turbidity*
(ex ante) impact assessment	*How does A affect B?*	*What is the likely impact of construction/ industrial activity on neighbouring land use*
Rights delineation	*How can the relevant property rights be established?*	*Can property boudaries be observed and delineated from EO data*
Policy formulation (intervention)	How can we best reduce the effects of E? How can we decrease/increase B? Which governance actors are best positioned to act? Which governance design is best suited to the problem?	Global environmental organization Local environmental organization Local government
Policy implementation	Has the selected intervention been implemented effectively? How can administrative costs be minimized?	Iceberg warning systems for secure route planning Early detection of food shortages and planning of relief operations Hurricane warning systems for planning of relief
Policy monitoring & enforcement	Did those assigned to reduce harm X do what they were supposed to do? How do we best track compliance with our program of E control? How do we best track compliance with (all sorts of) regulations? What indicators can be used for performance monitoring of implementation? How do we best track compliance with environmental treaties?	Agriculture Marine fisheries Marine oil spills Measures of urban form (landuse efficiency, density, etc)
Policy updating & refinement (evaluation)	Is the policy design adequate or has anyone come up with a better way (=knowledge breakthrough) to deal with the harm? Does this knowledge breakthrough change our assumptions about: - the existence of X - the decrease/increase of B - the distribution of property rights - intervention options - governance mechanisms?	Examples of changing assumptions?

Problem determination

The detection of the hole in the ozone layer is probably the best documented example of EO triggering policy development [19]. Concern about the detrimental effect of chlorofluorocarbons (CFCs) on the ozone layer [76] stimulated the US Congress to commission NASA to develop a sensor (TOMS) to monitor the state of the ozone layer. In 1986 NASA confirmed that deepening and enlargement of the size of the hole had indeed occurred since the late 1970s [109]. This confirmation, based on re-analysis of TOMS data, followed a report published in a paper one year earlier that alarmed the world by describing a steady decline of spring ozone concentrations over one British Antarctic Survey research station [31]. In response to these findings, the 1987 Montreal Protocol [117] prescribed a policy measure aimed at a 50% reduction of the use of hard CFCs, to be followed after four years by a complete ban on their use. Think for a moment about the fact that governments used command and control here, ending with a complete ban. Would you have done the same? Or would you prefer financial incentives? If yes, what, in your opinion, would have been the effect?

ozone and CFC

Policy implementation

In the Netherlands, remote sensing is used to support the implementation of a direct financial intervention that aims to tackle the effects of erosion that is weakening the effectiveness of measures to protect the coastal dune system against storms [115]. Airborne laser altimetry is used to estimate the volume of sand required for beach nourishment by measuring the elevation of the dry parts of the beach; sonar is used for the underwater parts. The Ministry of Public Works uses this information to direct companies that have been contracted to replenish the beaches. Maintenance of coastal defence lines is a task of the central government and EO contributes to the efficacy and efficiency of achieving this policy through assessment of the volume of sand required.

coastal defence

Policy monitoring

EO may also be used for policy monitoring in a various other ways, e.g. for environmental policy control in agriculture [16]. EU member states are obliged to check and report compliance with the regulations of the EU's Common Agricultural Policy (CAP). Information systems that combine the use of EO and land parcel-based information are used to cross-check whether subsidies received actually comply with EU regulations. Pedersen [89] gives a detailed description of the use of EO in crop-subsidy control in Denmark. While initially developed to detect subsidy fraud, the CAP system has recently been expanded to also include compliance with environmental directives.

precision farming

EO is also used to assess compliance with international treaties. For example, it has been recognized since the mid-1990s [42] that EO has potential for verifying compliance with international treaties [3, 63, 91] and increasingly it is being used by governments in implementing their commitments to these treaties, as well as checking compliance. In a similar context, EO is also being used to verify the compliance of contracting and non-contracting parties, as for the treaty to restrict the proliferation of nuclear weapons [82]. Recently, interest has increased in the application of EO to enforce the compliance of international environmental treaties, for example the Ramsar convention [99]. Although technically feasible, enforcing compliance with such treaties is not straightforward because they typically lack the legal framework needed for enforcement. Although EO alone cannot enforce compliance, it can be used to compile evidence of non-compliance. It can be used in governance processes that seek to achieve compliance by increasing pressure on the non-compliant parties to improve their performance.

Dealing with "wicked" problems: spatial planning as a form of governance

Many societal problems, such as how to achieve development, are "wicked" problems: this means that they are ill-defined, rely on political judgement for resolution, and defy solution in the sense that a well-defined engineering problem can be solved [98]. To meet the multiple challenges of development, as for example laid down in The Millennium Development Goals (MDGs), governments develop and implement policies, plans and projects that influence the spatial distribution of people, resources and activities at global, national, regional and local scales. Key sectors that address problems of food and water security, accessibility to infrastructure, shelter and social services, economic development, natural risks and biodiversity, and so on, all have a strong spatial dimension. At the same time, governments are faced with the problem of coordinating and integrating the spatial claims of different sectoral policies at different spatial administrative levels (national, regional or local). Spatial planning should balance these competing interests using legal, economic and communication policy instruments.

Spatial planning in its most formal sense is concerned with the problem of coordination or integration of the spatial dimension of different sectoral policies at a specific spatial administrative level (national, regional or local). This implies that spatial planning per definition addresses multiple objectives, (i.e. economic, social and environmental) and is multi-sectoral.

spatial planning

In spatial planning, different degrees and types of development and control mechanisms may be used. Whatever the governance approach and the level of public interference, spatial information is indispensable for supporting the various level of governments in the process of spatial planning for development. The aims of economic development, equity and environmental sustainability require a strong spatial vision that should be supported by various types of spatial information: this might include raw data at various scales: such as that provided by aerial photographs or satellite images of different kinds; topographic data at various scales describing the state of the planning area at a given time in the past; models describing past or perhaps future processes and conditions (e.g. economic, social, environmental); or perhaps take the form of scenarios or spatial development plans that represent a possible future world that should enjoy stakeholder support and may possibly be created through the integrated agency of these stakeholders.

1.5 A systems view of Earth processes: some examples

As it is impossible to be complete in representing the processes that act on the Earth, we need to focus on the most important and most illustrative cycles. The carbon and nitrogen geochemical cycles and the water cycle have been chosen as they impact on or are linked to global climate and change thereof. All of these cycles are important elements in the MDGs that are expressed in 21 quantifiable targets, which are to be measured by 60 indicators over the eight formulated goals:

- Goal 1: Eradicate extreme poverty and hunger;

- Goal 2: Achieve universal primary education;

- Goal 3: Promote gender equality and empower women;

- Goal 4: Reduce child mortality;

- Goal 5: Improve maternal health;

- Goal 6: Combat HIV/AIDS, malaria and other diseases;

- Goal 7: Ensure environmental sustainability; and

- Goal 8: Develop a global partnership for development.

For more information see `http://www.endpoverty2015.org/`.

1.5.1 The water cycle

hydroclimatic variables

To understand the role of the terrestrial hydrosphere–biosphere in the Earth's climatic system it is essential to be able to measure hydroclimate variables such as radiation, precipitation, evapotranspiration, soil moisture, cloud formation, water vapour, surface water and runoff, vegetation state, albedo and surface temperature. Figure 1.10 shows the major components of the water and energy cycles of the Earth system and some possible observed values. Such measurements are required to improve our understanding of global climate and its variability, both spatially and temporally. Additionally, such observations are essential for advancing our understanding of linkages between the terrestrial and atmospheric branches of the water cycle, and how these linkages may influence climate variability and predictability. To enhance prediction of the global water cycle an improved understanding of hydrological processes, their links with the energy cycle, as well as a sustained monitoring capability, are critical to mitigate water-related damage and sustainable human development. In many cases, the combination of space-based data and high-resolution *in situ* data provides a powerful tool for effectively addressing water management issues.

Water resources management, which can be considered as a product of governance processes, directly interferes with the natural water cycle through the building of dams, reservoirs, water-transfer systems and irrigation infrastructure that divert and redistribute part of the surface-water storages and fluxes. The water cycle is mainly driven and linked to the energy cycle through changes in the phases of water (between the liquid, vaporous and solid phases) and the transport of water (in addition to gravity and diffusion processes). Water cycle components can be observed with insitu sensors as well as airborne and satellite-borne radiation sensors. Processing and conversion of these radiation signals is necessary to retrieve data on the water cycle components.

water management

Water is an important aspect of governance, from the global through to the local level. The Millennium Declaration of the United Nations (UN) called on all members "to stop the unsustainable exploitation of water resources by developing water management strategies at the regional, national and local levels which promote both equitable access and adequate supplies". Improving water management can make a significant contribution to achieving most of the MDGs, especially those dealing with poverty, hunger and major diseases. The World Summit on Sustainable Development (WSSD) in 2002 emphasized in particular the importance of water and sanitation. The recommendations in the Johannesburg Plan of Implementation regarding water and sanitation are aimed at improving water resources management and scientific understanding of the water cycle through joint cooperation and research and, for this purpose, promoting knowledge sharing, providing capacity building and facilitating the transfer of technology. As mutually agreed, this includes remote sensing and satellite technologies, especially for developing countries and countries with economies in transition, and also includes supporting these countries in their efforts to monitor and assess the quantity and quality of water resources, for example by establishing and/or further developing national monitoring networks and water resources databases and by developing relevant national indicators. The Johannesburg Plan also adopted integrated water resources management (IWRM) as the overarching concept for addressing and solving water-related issues. As a result of the commitments made under

the Johannesburg Plan of Implementation, several global and regional initiatives have emerged.

According to a recent report prepared under the auspices of the Intergovernmental Panel on Climate Change (IPCC), "Observational records and climate projections provide abundant evidence that freshwater resources are vulnerable and have the potential to be strongly impacted by climate change, with wide-ranging consequences on human societies and ecosystems." Subject to climate change, the security of freshwater resources has emerged as a key societal problem. Floods, droughts, water scarcity, water usage, water quality, water and ecosystem interactions, and water and climate interactions are all issues of direct importance to human society. The only key to safeguarding the security of water resources is better water resources management. This in turn requires better understanding of the water cycle, water–climate and water–ecosystem interactions, and effective governance systems that can implement the actions that have been agreed to.
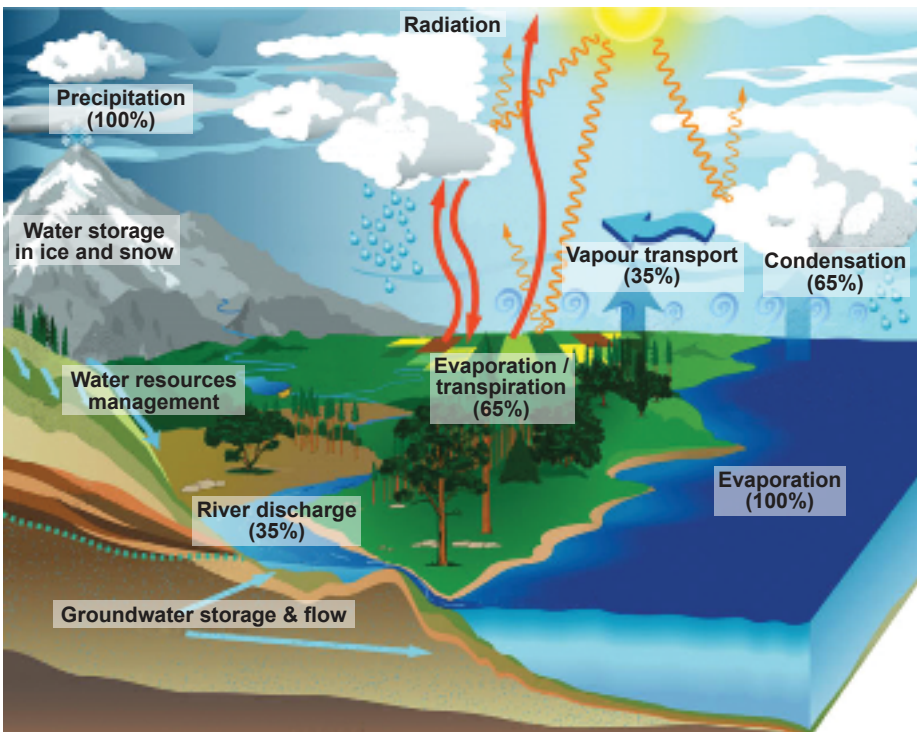
Figure 1.10
Global water and energy cycle of the Earth system and some possible observed values. Courtesy NASA.

1.5.2 The carbon (C) cycle

Carbon (C) is found in solid form in most organisms, as well as occurring in gaseous form as carbon dioxide. The carbon cycle (Figure 1.11) can be divided into a geological cycle that operates on time scales in the order of millions of years and a biological cycle that operates on a time scale of days to thousands of years. The geological component of the carbon cycle is part of the rock cycle in the processes of weathering and dissolution, precipitation of minerals, and burial and subduction. In the ocean, within the context of geological processes, carbon is stored in the form of calcium carbonates, which are present in various reef-forming organisms. Oceanic volcanism—spreading lava across sea floors—is responsible for generating new crust, while in subduction zones the crust is recycled. In both processes, carbon and associated gases are formed

and subsequently destroyed. The biological carbon cycle transfers carbon between land, ocean and atmosphere through processes of photosynthesis and respiration.

Plants take in carbon dioxide (CO_2) from the atmosphere during photosynthesis and release CO_2 back into the atmosphere during respiration:

Respiration: $C_6H_{12}O_6$ (organic matter) + $6O_2$ = $6CO_2$ + $6H_2O$ + energy

Photosynthesis: energy (sunlight) + $6CO_2$ + $6H_2O$ = $C_6H_{12}O_6$ + $6O_2$

Through photosynthesis, green plants use solar energy to turn atmospheric carbon dioxide into carbohydrates (sugars). Plants and animals use these carbohydrates (and other products derived from them) through a process called respiration—the reverse of photosynthesis. Respiration releases the energy contained in the sugars for metabolic activity and changes carbohydrate *fuel* back into carbon dioxide, which is in turn released back into the atmosphere. In addition to natural carbon cycles, additional amounts of carbon dioxide are introduced into the system by human activity, bringing with it adverse effects on climate and the potential of the Earth to smother itself. To re-iterate, the negative effects of CO_2 enrichment in the atmosphere by human activity are not fully understood, nor are we presently capable of mitigating these effects.

Greenhouse effects and the carbon cycle, in particular carbon emissions and carbon sequestration, are at the heart of climate change, one of the most pressing problems the Earth is currently facing. The Intergovernmental Panel on Climate Change (IPCC), an inter(national)disciplinary consortium of climate experts, has produced various reports and inventories to asses the problem; these led to the agreement known as the Kyoto Protocol, aimed at combatting global CO_2 emissions. Accurate quantification of the various components of the carbon cycle represents a core requirement for the assessment, monitoring, modelling, and mitigation of adverse climate effects and, in the end, sustainability of livelihoods in many parts of the globe. The latter also requires a governance process for the identification, analysis and the development of policy instruments in order to deal with the impacts of foreseeable changes in the carbon cycle. Within the carbon cycle, forestry (in the broadest sense) is the principal area of interest for emissions (sources) and sequestration (sinks). Afforestation, reafforestation and deforestation are the current Kyoto focal points, but sustainable forest management, including certification and the assessment and prevention of forest degradation, may well be considered in the post-Kyoto period. Due to size, inaccessibility of the forest resources, and international requirements for a uniform methodology, the quantification of the components of the carbon cycle in both space and time depends heavily on remote sensing, GIS modelling and related statistical tools. Nevertheless, there are significant gaps in knowledge in these fields. Still more knowledge gaps exist when facing the post-Kyoto situation with respect to the assessment and monitoring of forest degradation and land cover change, in general, and their relationships with biomass and carbon. To assess the likely impacts of changes in the carbon cycle, and, thus, the likely effect on climate—and especially for local communities, there is also a great need for "ground truthing" of climate scenarios and macro data. Areas of uncertainty include the identification and quantification of land-based sources and sinks; the tracking of progress towards achieving emission targets; the formulation of policy and management guidelines for proper vegetation and land management; and the assessment of relationships between sustainable forest management and biomass sequestration, as well as the relationship biomass–forest degradation.

1.5.3 The Nitrogen (N) cycle

Nitrogen (N)is an essential component for the growth of organisms and is one of the main constituents of DNA, protein and leaf pigments. Plants and animals take nitrogen up in the form of ammonium (NH^{4+}), nitrate (NO^{3-}) or organic nitrogen (e.g.

Atmosphere 750

CO₂

5.5

60

121.3

60

Vegetation 610
1.6

0.5

Fossil Fuels &
Cement Production
4,000

92

90

Soils
1,580

50

Surface Ocean 1,020

40

Marine Biota 3

6

4

Deep Ocean
18,100

Dissolved Organic
Carbon <700

6

0.2

Sediments 150

**Storage in GIC
Fluxes in GIC/yr**
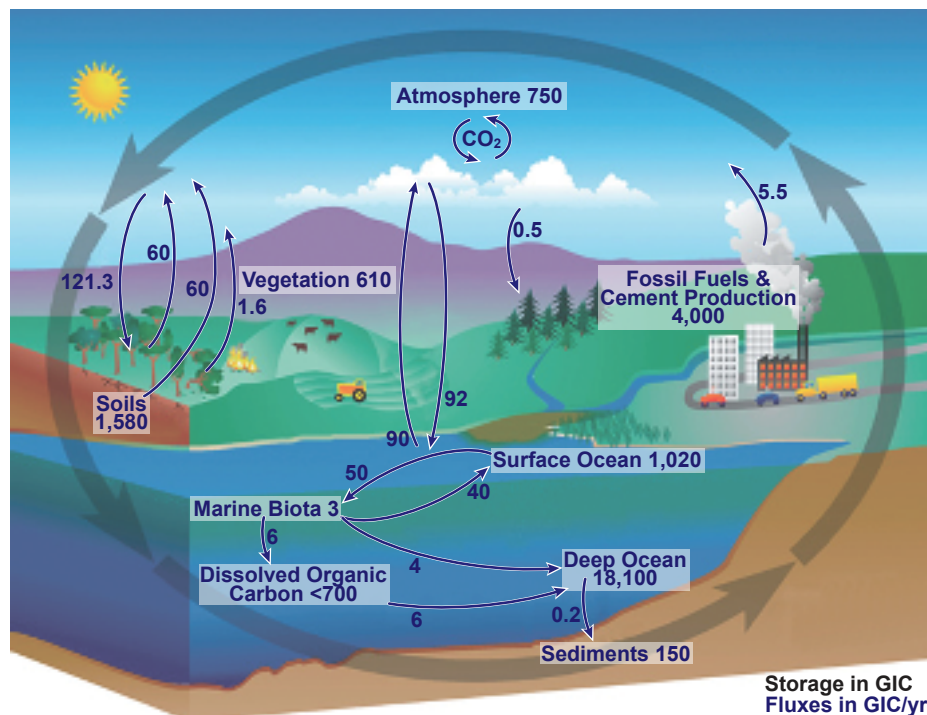
Figure 1.11
The carbon cycle. Courtesy
NASA/JPL-Caltech.

NH_3 2CO), which means that for nitrogen to be of use to living organisms it needs to be transferred to these solid states. The nitrogen cycle describes the movement of nitrogen between and within the atmosphere, biosphere and geosphere. Together with the carbon cycle, the nitrogen cycle (Figure 1.12) is one of the most important geochemical cycle on Earth. It actually consists of a number of processes, of which fixation, uptake, mineralization, and nitrification are the most important ones to consider. Fixation is the process by which nitrogen is converted to ammonium by bacteria through metabolic processes. Metabolic processes are chemical reactions that allow organisms to grow and mutate and as a result produce or change element stages. The ammonia produced by nitrogen-fixing bacteria is incorporated into protein and other organic nitrogen compounds, either by a host plant, the bacteria itself, or other organisms present in the soil. This process is referred to as uptake of nitrogen. After nitrogen is incorporated into organic matter, it is often converted back into inorganic nitrogen by a process called nitrogen mineralization. When organisms die, bacteria consume the organic matter, which leads to decomposition of material and formation of ammonium. In the presence of oxygen, some of the available ammonium is converted to nitrate by the bacteria (nitrification), resulting in a net gain of energy. Humans have influenced (and thus altered) the nitrogen cycle by introducing synthetic nitrogen fertilizers to increase agricultural productivity. These fertilizers are taken up by crops to accelerate the metabolic processes, with the adverse effect that some parts of this end up in groundwater and surface water, leading to nutrient enrichment that causes algal blooms and disruption of natural ecosystems. Reactive nitrogen also ends up in the atmosphere as smog, thus causing long-term changes to global climate. Some of the nitrogen ends up in "acid rain", which is a popular term describing precipitation that contains carbon- and nitrogen-related particles resulting from human emissions—typically from industrial activities—of sulfur and nitrogen compounds into the atmosphere. These processes destabilize the natural balance of ecosystems, although the

long-term consequences still remain relatively unknown.
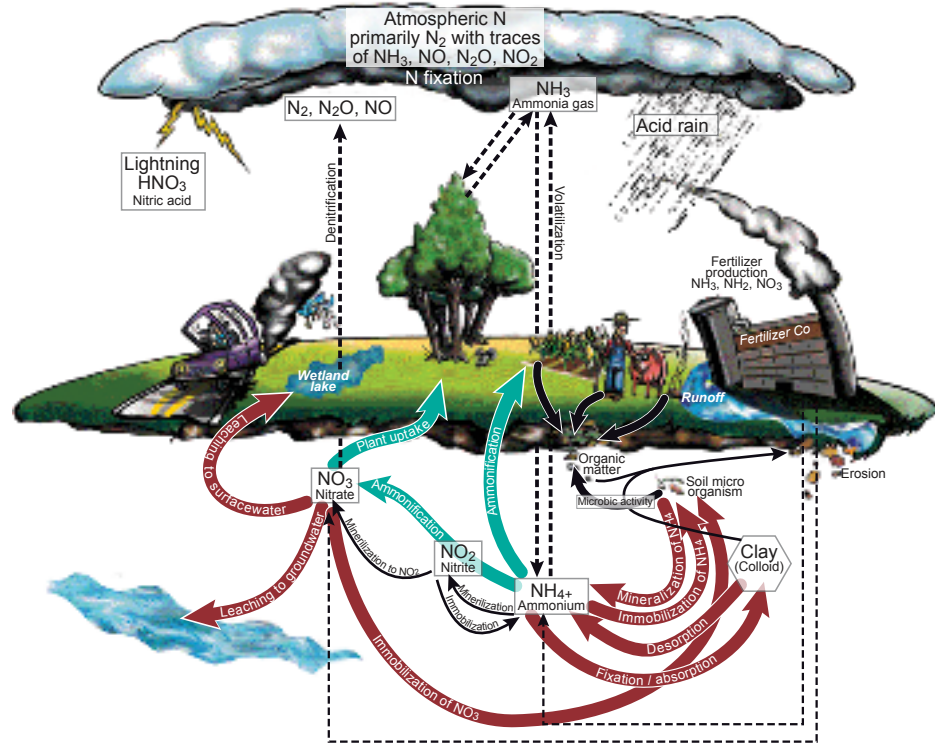
Figure 1.12
The nitrogen cycle. Courtesy NASA/JPL-Caltech.

1.5.4 Urbanization

Urbanization refers to an increase in the proportion of a population living in settlements that are considered to be urban. According to United Nations statistics, 2008 was marked by a global urbanization rate of 50% (see Figure 1.13). This milestone is of particular importance for a variety of reasons. Urbanization is generally linked to and accompanied by economic transformations associated with industrialization and modernization and, owing to the economies of scale offered by cities, urban areas provide the basis for high productivity and high economic growth rates that far exceed the performance of rural areas. Urbanization is therefore a key element of development policies and poverty reduction strategies.
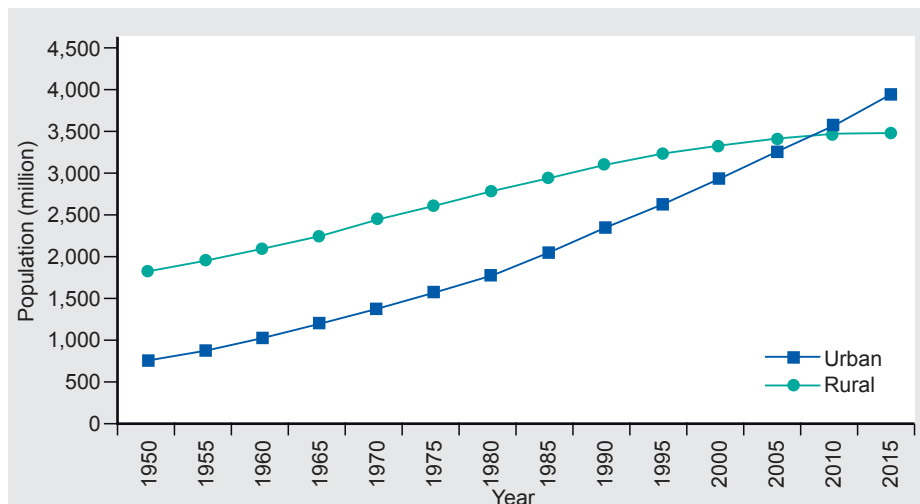
Figure 1.13
Growth in the world's rural and urban population, 1950–2000, with projections up to 2015.

Annez and Buckley [1] also point to three major concerns that have been frequently expressed about the effects and desirability of urbanization. First, a high rate of rural–urban migration may be seen as unmanageable and undesirable as it leads to excessive demands for serviced land, housing, social services, etc. This may result in the proliferation of informal settlements and slums, a commonly-observed phenomenon in many cities of developing countries. Second, there is a long-lasting belief that urbanization in developing countries is economically counter-productive, because its large informal sector includes massive amounts of hidden unemployment. The third criticism is that of pro-urban bias, which refers to the notion that government policies tends to favour investment in urban areas rather than in rural areas, thereby contributing to the relative attractiveness of urban areas, fuelling increased rural–urban migration. More recent thinking tends, however, to emphasize the interrelationships between urban and rural areas, and evidence for the benefits of creating and supporting the linkages between urban areas and their rural hinterlands continues to grow.

Another major issue that presently attracts much attention is the link between urbanization and climate change, and the effects that climate change is having on the lives of people around the world. Urban areas are major contributors to global green house gas (GHG) production. The International Panel on Climate Change (IPCC) reported on the sources of GHG emissions by sector in 2004 (see Figure 1.14). The high energy demands associated with urban activities are the major contributor to GHG emissions, thus the role of cities in the mitigation of climate change is increasingly important. Moreover, many cities need to adapt to the new environmental conditions associated with climate change: cities located close to the coastlines need to consider the effects of

sea-level rise; cities also need to consider the impact of changing patterns of rainfall on storm-water runoff and flooding, or on drought; rising global temperatures may exacerbate urban-heat-island effects, which contribute to higher heat-induced morbidity and mortality rates. Such changes have implications for urban planning and management, and therefore require policy change at national and local levels. In addition, the effects of climate change will also be felt in rural areas and could potentially lead to environmental changes that will contribute to further increases in urbanization.
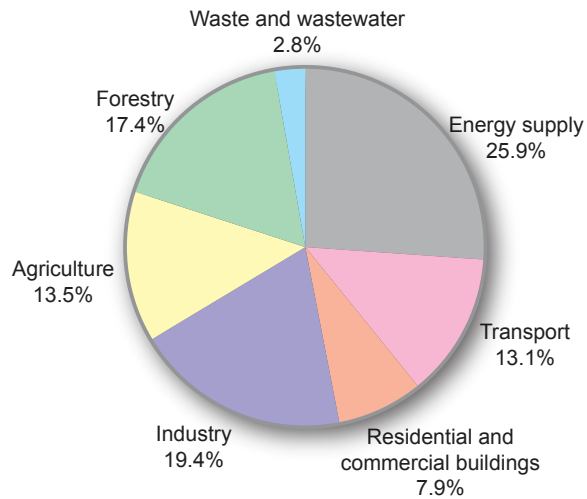
Figure 1.14
Global GHG emissions by sector in 2004.
Source: IPCC 4.

Cities and their urban regions are highly complex and dynamic entities. The observable physical changes in the built environment do not fully reflect the nature of their complexity and dynamics. For example, important socio-economic processes associated with intra-urban migration such as gentrification (the gradual displacement of low-income households in a neighbourhood by high-income households through property acquisition) or filtering (the opposite process of low-income households displacing higher-income households) are important phenomena from an urban-management perspective, although they may not be directly measurable through EO. In addition to such socio-economic processes, urban planners and managers are very much concerned with other processes for which spatial information can provide useful and sometimes critically important information. The search for solutions to issues related to urban land use development, infrastructure, traffic and transport, the environment, and governance can benefit from the use of geoinformation technology, although a good understanding of the local context in which urban development is taking place is always necessary. For example, urban areas in the Netherlands have grown quite considerably over the last 20 years, but, unlike the situation in many developing countries, this expansion has not been driven by urban population growth due to migration. Rather the growth of urban areas can be attributed to increased wealth, increases in per capita demand for space and improvements in mobility.

What is needed are planning approaches that are relevant to the cities of the future, that are grounded on a view of cities and human-settlement networks as complex, dynamic entities that interact with their environs (i.e. the various natural spheres and subsystems of system Earth) in a variety of ways that are significant for the life and well-being of current and future world populations. This means that traditional concerns of urban planners and designers with urban forms and spatial development patterns should be married to a deeper analysis and understanding of the environmental impacts of urban development in the short, medium and long term. Fundamental con-

cerns about the sustainability of cities and human settlements must develop ways of reducing the consumption of all non-renewable resources (such as land, water, air and minerals) and protecting them from depletion and degradation through inappropriate waste disposal. Urban planning and management processes will therefore need to include methods for analysis and assessment that consider the flow of materials, energy and waste associated with urban development as explicit components of the planning development and implementation processes.

Urban planning and management can, for example, play a role in reducing GHG emissions by adopting land use policies that promote more compact urban forms characterized by land use mixing and improved public transport systems. EO and GIS data can be used in the planning and design process to develop and analyse scenarios of how this might be done. Moreover, the analysis of the effects of heat islands and energy loss from buildings, as well as the assessment of the influence of building form, orientation and juxtaposition on the liveability of urban spaces, is an area in which GI Science has potential given the likelihood of significant warming in many parts of the world. What this boils down to is the development and adoption of approaches to urban development that are grounded on the best available evidence and understanding of how cities and towns function and interact with their environs. These approaches must also support the kind of multi-level, multi-actor governance approaches to development that are now seen to be fundamental to successful planning implementation. The techniques and tools of GI Science are already being used to support collaborative planning and decision-making, but further research, development, testing and dissemination is still required.

1.6 Concluding remarks

Clearly the world is not ideal place. System Earth is ridden with complexity and although our knowledge of its workings has perhaps never been so great, our lack of knowledge and understanding still confronts us and confounds our best-made plans daily.

In many developing countries the limits to knowledge and planning are very obvious. Data, information and knowledge are more likely to be incomplete or of poor quality, and the economic, human and technical resources needed for planning and development may also be inadequate. But even under such highly constrained circumstances, the development of models in interaction with integrated spatial plans will be required. Effective use of scientific tools and GISs, and Earth Observation techniques can help to establish a policy cycle in which our knowledge of system Earth is developed, managed and exploited for the purpose of sustainable development.

We have seen that a great variety of information about system Earth is needed in the governance processes for system Earth. Some of this information can be derived from EO and combined with that from other sources. To understand the processes affecting system Earth, we first need to understand what the main issues, concepts and relationships are. We also need to be able to properly monitor the underlying processes and break them down into key *observables* that can be monitored over long periods of time. Some of the "big questions" about system Earth that have been identified are:

observables

- How is the global Earth system changing?

- What are the primary forces of the Earth system?

- How does the Earth system respond to natural and human-induced changes?

- What are the consequences of change in the Earth system for human civilization?

- How should scenarios for local environmental change be considered in a local context?

- How might a community, city, or region best adapt to changing environmental conditions?

But many of the geographic data sets needed often have shortcomings, such as:

- The data may not be available in a format readable by the user.

- The benefit gained from using the data may not weigh up against the cost involved in acquiring image data.

- Delays in data availability (e.g. real time data needed in cases of natural disasters) may prevent timely use of EO technology.

- Spatial and temporal coverage may not be optimal for the thematic application that is being catered for.

- It may not be possible to combine observations from different data agencies because there are no agreed standards.

- Data discontinuities may occur owing to gaps in data acquisition.

- Global level data may not be readily scalable to local areas, despite the need at that level for local policy development and decision-making.

One of the most common pitfalls encountered when using geographic data is that there has been inadequate involvement of stakeholders in the definition of the products and information requirements, and, linked to this, there are no exhaustive agreements on data standards, so interoperability is not guaranteed. One solution to this problem is a closed-chain approach in geoinformation provision. Such a chain starts with a stakeholder-defined question, which can be driven by issues of societal, economic, scientific or political relevance. Both producers and users of geoinformation should have a common perception of the problems at hand and the need for information. Some of the data needed can be acquired through EO acquisition (satellite-based and airborne,) while other data may be drawn from existing archives or acquired by field measurement. In the examples in Chapter 12, further attention is given to the use and role of stakeholders in different settings.

The meteorological sector is an example of a user community that for years has been able to sustain a steady flow of relevant products from EO and field data that has driven innovations in weather forecasting. The sector has delivered new products such as rainfall radar and UV index mapping, to mention just two. Similarly,in the wake of the climate debate, the atmospheric chemistry community has been successful in closing the EO chain and securing lengthy time series of data sets on, for example, water vapour, trace gases and aerosols. These communities have in common that they are well organized and that they have clear demands and well-specified products.

	Disasters	Health	Energy	Climate	Water	Weather	Ecosystems	Agriculture	Biodiversity
Disasters		Early warning of likely injury or loss of life	Risk to energy infrastructure	Volcanic eruptions Fire emissions	Risk to water infrastructure	Volcanic eruptions Fire emissions	Ecosystem disturbance Ocean pollution	Risk to crops and livestock, farming land and pastureland	Population disturbance Ocean pollution
Health	Vulnerability of exposed population				Sources of pathogens		Spreading of endemic disease. Distribution of vectors		Distribution of vectors
Energy	Loss of energy services	Heating and cooling. Pollutants of air, water and by heath		Emission of greenhouse gasses	Pollutants Energy for water services	Air quality	Mining of fossil fuel. Fuel wood and peet	Energy for agriculture	Disturbance due to energy activities
Climate	Projections of climate change	Epidemiology of vectorborn diseases. Other climate sensitive diseases	Wind and solar energy potential		Expected level and variability of water balance	Climate change, anomaly identification	Climate and disturbance data, climate change	Climate data Climate change Desertification	Climate and disturbance data
Water	Risk of floods and draught	Water borne pathogens, pollutants. Availability of save water for drinking.	Hydroelectric potential	Water cycle and energy budgets		Fluxes of energy and water, surface boundary conditions	Water amount Water quality	Availability and suitability for irrigation	Water amount Water quality
Weather	Extreme weather forecasts	Extreme events. Pollen allergy	Changes in demand. Risk to energy infrastructure	Long term statistics	Precipitation Evaporation		Disturbance Droughts Productivity effects Fire risk	Rainfall Frosts Temperature	Meteorological data for survival of species
Ecosystems	Vulnerability to flood and landslides. Fire risk. Recovery after disaster	Ecological condition for outbreak of disease vector. Red tide. Sand storms.	Biomass for energy. Carbon cycle	Water budget Carbon cycle Trace gases Albedo	Land cover	Land cover		Weeds Pests Diseases Rangelands Wild fisheries	Habitat distribution and fragmentation Ecosystem function
Agriculture	Food insecurity	Nutrition. Food savety. Vector born diseases.	Biomass fuels	Greenhouse gas emissions Land cover and Land use	Use of irrigation Water pollution Water storage	Land cover Land use	Loss of area Eutrophication Fertilizer application Chemical application		Agricultural biodiversity Distribution of GMO. Trade statistics on alien organisms
Biodiversity	Locust plagues	Distribution of medical plants, microbes and disease vectors. Spread of endemic diseases.			Bioindicators of water quality	Abundance and distribution of organisms Phenology	Abundance and distribution of organisms Phenology Prediction of change in distribution Invasive species	Agricultural biodiversity	

In contrast, the terrestrial "remote sensing" community is quite heterogeneous, less well organized and has a wide range of demands in terms of data products. The overview of system Earth and its subsystems presented in Table 1.2, which features the key cyclic processes, represents an attempt to produce an exhaustive list of essential parameters for Earth Observation.

It is impossible to produce an exhaustive list of interrelated variables or observables that would allow us to monitor, model and understand system Earth in all its facets. Many key observables have been defined by the Group on Geoinformation (GEO; `http://www.earthobservations.org/`) in their overview of information provision for nine interrelated areas of societal relevance: disasters, health, energy, climate, water, weather, ecosystems, agriculture and biodiversity. Figure 1.15 shows some of the major connections between these nine areas of interest. For a complete list of the variables, the reader is referred to GEOSS's 10-year implementation plan. In the following chapters of this book, you will learn more about how to collect and use these and other data in the governance of system Earth.

Chapter 2

Physics

Wan Bakx
Klaus Tempfli
Valentyn Tolpekin
Tsehaie Woldai

Introduction

Geospatial Data Acquisition (GDA) challenges us to make choices: on which one of the many sensors available should the agronomist rely for accurate yield predictions? If he or she chooses a sensor producing several images, such as a multispectral scanner, which image or which combination of images to use? How to properly process sensor recordings to increase the chances of a correct interpretation? When interpreting a colour image, what causes the sensation *red*? Instead of writing a thick book of recipes to answer such questions for every application, we can better review the physics of RS. Understanding the basics of electromagnetic (EM) radiation will help you in making more profound choices and enable you to deal with sensors of the future.

A standard photograph is an image of an object or scene that very closely resembles direct sensing with our eyes. The sensation of colour is caused by EM radiation. Red, green and blue relate to forms of radiation that we commonly refer to as light. *Light* is EM radiation that is visible to the human eye. As we are interested in Earth Observation, our light source is the Sun. The Sun emits light, the Earth's surface features reflect light, and the photosensitive cells (cones and rods) in our eyes detect light. When we look at a photograph, it is the light reflected from the photograph that allows us to interpret the photograph. Light is not the only form of radiation from the Sun and other bodies. The sensation *warm*, for example, is is the result of thermal emissions. Another type of emissions, ultraviolet (UV) radiation, triggers our body to generate vitamin D and also produces a suntan.

This chapter explains the basic characteristics of EM radiation, its sources and what we call the EM spectrum, the influence of the atmosphere on EM radiation, interactions of EM radiation with the Earth's surface, and the basic principles of sensing EM radiation and generic properties of sensors.

2.1 Waves and photons

EM radiation can be modelled in two ways: by waves, or by radiant particles called photons. The first publications on the wave theory date back to the 17th century. According to the wave theory, light travels in a straight line (unless there are external influences) with its physical properties changing in a wave-like fashion. Light waves have two oscillating components: an electric field and a magnetic field. We refer, therefore, in this context to electromagnetic waves. The two components interact— an instance of a positive electric field coincides with a moment of negative magnetic field (Figure 2.1). The wave behaviour of light is common to all forms of EM radiation. All EM waves travels at the speed of light, which is approximately equal to 2.998×10^8 m s^{-1}. This is fast, but the distances in space are literally astronomical: it takes eight minutes for the sunlight to reach the Earth, thus when we see, a sunrise, for example, the light particles actually left the Sun that much earlier. Because they travel in a straight line, we use the notion of light rays in optics.
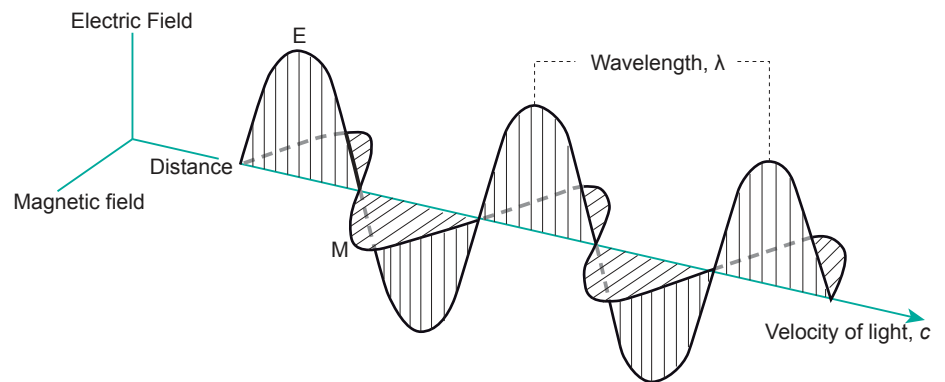
Figure 2.1
The two oscillating components of EM radiation: an electric and a magnetic field.

A sine wave can be described as:

$$e = \alpha sin\left(\frac{2\pi}{\lambda}x + \varphi\right). \tag{2.1}$$

where α is the amplitude of the wave, φ is the phase (it depends on time) and λ is the *wavelength*. The wavelength is a differentiating property of the various types of EM radiation and is usually measured in micrometres (1 μm $= 10^{-6}$ m). Blue light is EM radiation with a wavelength of around 0.45 μm. Red light, at the other end of the colour spectrum of a rainbow, has a wavelength of around 0.65 μm (Figure 2.2). Electromagnetic radiation outside the range 0.38–0.76 μm is not visible to the human eye.
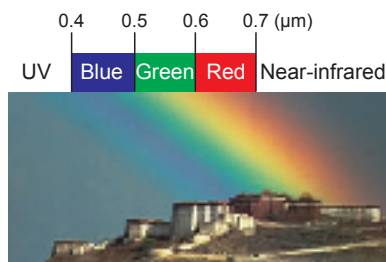
Figure 2.2
The spectrum of light.

We call the amount of time needed by an EM wave to complete one cycle the *period* of

the wave. The reciprocal of the period is called the *frequency* of the wave. Thus, the frequency ν is the number of cycles of the wave that occur in one second. We usually measure frequency in hertz (1 Hz = 1 cycle s^{-1}). Since the speed of light c is constant, the relationship between wavelength and frequency is:

$$c = \lambda \times \nu. \tag{2.2}$$

Obviously, a short wavelength implies a high frequency, while long wavelengths are equivalent to low frequencies. Blue light has a higher frequency than red light (Figure 2.3).

Short wavelength   Long wavelength

High frequency   Low frequency
High energy   Low energy
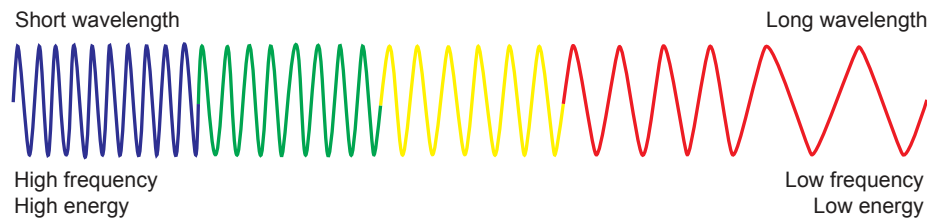
Figure 2.3
Relationship between wavelength, frequency and energy.

Although wave theory provides a good explanation for many EM radiation phenomena, for some purposes we can better rely on particle theory, which explains EM radiation in terms of photons. We take this approach when quantifying the radiation detected by a multispectral sensor (see Section 2.6). The amount of energy carried by a photon of a specific wavelength is:

energy of a photon

$$Q = h \times \nu = h \times \frac{c}{\lambda}, \tag{2.3}$$

where Q is the energy of a photon measured in joules (J) and h is Planck's constant ($h \approx 6.626 \times 10^{-34}$ J s).

The energy carried by a single photon of light is just sufficient to excite a single molecule of a photosensitive cell of the human eye, thus contributing to vision. It follows from Equation 2.3 that long-wavelength radiation has a low level of energy while short-wavelength radiation has a high level. Blue light has more energy than red light (Figure 2.3). EM radiation beyond violet light is progressively more dangerous to our body as its frequency increases. UV radiation can already be harmful to our eyes, so we wear sunglasses to protect them. An important consequence of Formula 2.3 for RS is that it is more difficult to detect radiation of longer wavelengths than radiation of shorter wavelengths.

2.2   Sources of EM radiation

All matter with a temperature above absolute zero emits EM radiation because of molecular agitation. *Planck's law of radiation* describes the amount of emitted radiation per unit of solid angle in terms of the wavelength and the object's temperature:

Planck

$$L(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k T}} - 1}, \tag{2.4}$$

where h is the Planck's constant, $k \approx 1.38 \times 10^{-23}$ J K^{-1} is the Boltzmann constant, λ is the wavelength (m), c is the speed of light and T is the absolute temperature (K). $L(\lambda, T)$ is called the spectral radiance.

We can use different measures to quantify radiation. The amount of radiative energy is commonly expressed in joules (J). We may, however, be interested in the radiative energy per unit of time, called the *radiant power*. We measure the power in watts (W = J s^{-1}). *Radiant emittance* is the power emitted from a surface; it is measured in watts per square metre (W m^{-2}). *Spectral radiant emittance* characterizes the radiant emittance per wavelength; it is measured in W m^{-2} μm^{-1} (this is the unit used in Figure 2.4). *Radiance* is another quantity frequently used in RS. It is the radiometric quantity that describes the amount of radiative energy being emitted or reflected in a specific direction per unit of projected area per unit of solid angle and per unit of time. Radiance is usually expressed in W sr^{-1} m^{-2} (sr is steradian, unit of solid angle). *Spectral radiance* used in Equation 2.4 is radiance per wavelength and is measured in W sr^{-1} m^{-2} μm^{-1}. *Irradiance* is the amount of incident radiation on a surface per unit of area and per unit of time. Irradiance is usually expressed in W m^{-2}.

Planck's law of radiation is only applicable to black bodies. A black body is an idealized object with assumed extreme properties that helps us when explaining EM radiation. A black body absorbs 100% of incident EM radiation; it does not reflect anything and thus appears perfectly black. Because of its perfect absorptivity, a black body emits EM radiation at every wavelength (Figure 2.4). The radiation emitted by a black body is called black-body radiation. Real objects can re-emit some 80 to 98% of the radiation received. The emitting ability of real objects is expressed as a dimensionless ratio called emissivity $\epsilon(\lambda)$ (with values between 0 and 1). The *emissivity* of a material depends on the wavelength; it specifies how well a real body made of that material emits radiation as compared to a black body.

The Sun behaves similarly to a black body. It is a prime source of the EM radiation that plays a role in Earth Observation, but it is not the only source. The global mean temperature of the Earth's surface is 288 K and over a finite period the temperatures of objects on the Earth rarely deviate much from this mean. The surface features of the Earth therefore emit EM radiation. Solar radiation constantly replenishes the energy that the Earth radiates into space. The Sun's temperature is about 6000 K. Planck's law of radiation is illustrated in Figure 2.4 for the approximate temperature of the Sun (about 6000 K) and the ambient temperature of the Earth's surface (288 K). The figure shows that for very hot surfaces (e.g. the Sun), spectral emittance of a black body peaks at short wavelengths. For colder surfaces, such as the Earth, spectral emittance peaks at longer wavelengths. This behaviour is described by *Wien's displacement law*:

$$\lambda_{max} = \frac{b}{T}, \tag{2.5}$$

where λ_{max} is the wavelength of the radiation maximum (μm), T is the temperature (K) and $b \approx 2898$ μm K is a physical constant.

We can use Wien's law to predict the position of the peak of the black-body curve if we know the temperature of the emitting object. The temperature of the black body determines the most prominent wavelength of black-body radiation. At room temperature, black bodies emit predominantly infrared radiation. When a black body is heated beyond 4450 K (approximately 4700 °C) emission of light becomes dominant, from red, through orange, yellow, and cyan, (at 6000 K) to blue, beyond which the emitted energy includes increasing amounts of ultraviolet radiation. At 6000 K a black body emits radiation of all visible wavelengths in approximately equal amounts, creating the sensation of white to us. Higher temperatures correspond to a greater contribution of radiation of shorter wavelengths.

The following description illustrates the physics of what we see when a blacksmith heats a piece of iron or what we observe when looking at a candle. The flame appears

light-blue at the outer edge of its core; there the flame is hottest, with a temperature of 1670 K. The centre, with a temperature of 1070 K, appears orange. More generally, flames may burn with different colours (depending on the material being burnt, the surrounding temperature and the amount of oxygen present) and accordingly have different temperatures (in the range of 600 °C to 1400 °C). Colour tells us something about temperature. We can use colour, for example, to estimate the temperature of a lava flow from a safe distance. More generally, if we can build sensors that allow us to detect and quantify EM radiation of different wavelengths (also outside the visible range), we can use RS recordings to estimate the temperature of objects. You may also notice from the black-body radiation curves (Figure 2.4) that the intensity of EM radiation increases with increasing temperature; the total radiant emittance at a certain temperature is the area under the spectral emittance curve.
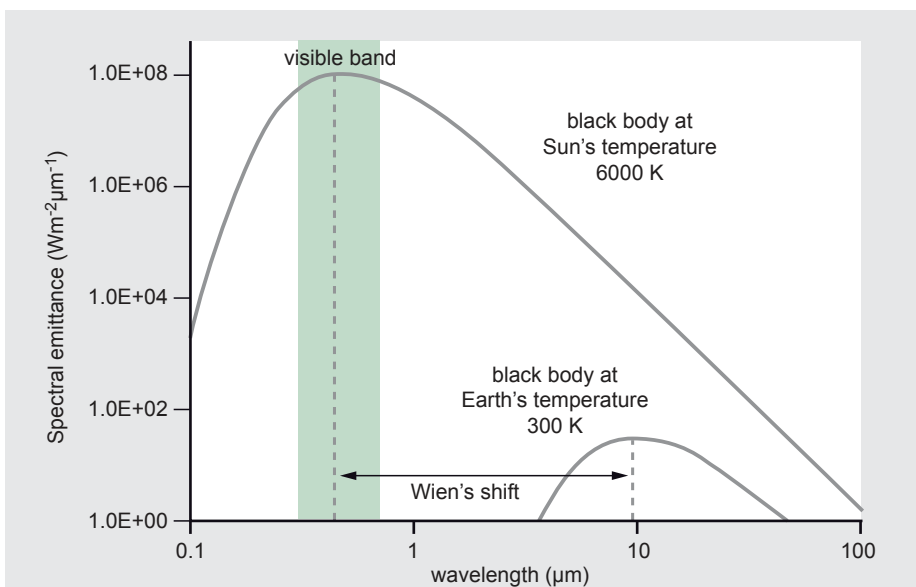
Figure 2.4
Illustration of Planck's law of radiation for the Sun (6000 K) and for the average surface temperature (300 K) the Earth. Note the logarithmic scale for both x- and y-axes. The broken lines mark the wavelength of the emission maxima for the two temperatures.

If you were interested in monitoring forest fires, which typically burn at 1000 K, you could immediately turn to wavelength bands around 2.9 µm, where the radiation maximum for those fires is to be expected. For ordinary land surface temperatures of around 300 K, wavelengths from 8 to 14 µm are most useful.

You can probably now understand why reflectance remote sensing (i.e. based on reflected sunlight) uses short wavelengths in the visible and short-wave infrared, and thermal remote sensing (based on emitted Earth radiation) uses the longer wavelengths in the range 3–14 µm. Figure 2.4 also shows that the total energy (integrated area under the curve) is considerably higher for the Sun than for the cooler Earth's surface. This relationship between surface temperature and total amount of radiation is described by the *Stefan-Boltzmann law*.

Stefan-Boltzmann law

$$M = \sigma T^4, \qquad (2.6)$$

where M is the total radiant emittance (W m^{-2}), σ is the *Stefan-Boltzmann constant* ($\sigma \approx 5.6697 \times 10^{-8}$ (W m^{-2} K^{-4}), and T is the temperature in K.

The Stefan-Boltzmann law states that colder objects emit only small amounts of EM radiation. Wien's displacement law predicts that the peak of the radiation distribution will shift to longer wavelengths as the object gets colder. In Section 2.1 you will have

learnt that photons at long wavelengths have less energy than those at short wavelengths. Hence, in thermal RS we are dealing with a small amount of low energy photons, which makes their detection difficult. As a consequence of that, we often have to reduce spatial or spectral resolution when acquiring thermal data, to guarantee an acceptable signal-to-noise ratio.

2.3 Electromagnetic spectrum

We call the total range of wavelengths of EM radiation the *EM spectrum*. Figure 2.2 illustrates the spectrum of visible light; Figure 2.5 illustrates the wider range of EM spectrum. We refer to the different portions of the spectrum by name: gamma rays, X-rays, UV radiation, visible radiation (light), infrared radiation, microwaves, and radio waves. Each of these named portions represents a range of wavelengths, not one specific wavelength. The EM spectrum is continuous and does not have any clear-cut class boundaries.
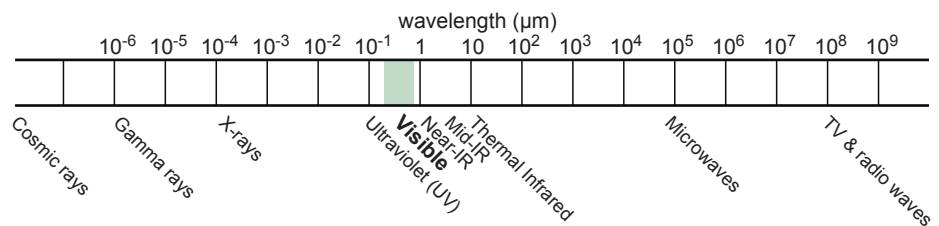
Figure 2.5
The EM spectrum.

Different portions of the spectrum have differing relevance for Earth Observation, both in the type of information that we can gather and the volume of geospatial data acquisition (GDA). The majority of GDA is accomplished by sensing in the visible and infrared range. The UV portion covers the shortest wavelengths that are of practical use for Earth Observation. UV radiation can reveal some properties of minerals and the atmosphere. Microwaves are at the other end of the useful range for Earth Observation; they can, among other things, provide information about surface roughness and the moisture content of soils.

The "visible portion" of the spectrum, with wavelengths producing colour, is only a very small fraction of the entire EM wavelength range. We call objects "green" when they reflect predominately EM radiation of wavelengths around 0.54 µm. The intensity of solar radiation has its maximum around this wavelength (see Figure 2.8) and the sensitivity of our eyes is peaked at green-yellow. We know that colour effects our emotions and we usually experience green sceneries as pleasant. We use colour to distinguish between objects and we can use it to estimate temperature. We also use colour to visualize EM radiation we cannot see directly. Section 5.1 elaborates how we can "produce colour" by adequately "mixing" the three primary colours red, green and blue.

light and colour

Radiation beyond red light, with larger wavelengths in the spectrum, is referred to as infrared (IR). We can distinguish vegetation types and the stress state of plants by analysing *near-infrared* (and *mid-infrared*) radiation—this works much better than trying to do so by colour. For example, deciduous trees reflect more near-infrared (NIR) radiation than conifers do, so they show up brighter on photographic film that is sensitive to infrared. Dense green vegetation has a high reflectance in the NIR range, which decreases with increasing damage caused by plant disease (see also Section 2.5.1). Mid-IR is also referred to as short-wave infrared (SWIR). SWIR sensors are used to monitor surface features at night.

near-infrared, short-wave infrared

Infrared radiation with a wavelength longer than 3 μm is termed thermal infrared (TIR) because it produces the sensation of "heat". Near-IR and mid-IR do not produce a sensation of something being hot. Thermal emissions of the Earth's surface (288 K) have a peak wavelength of 10 μm (see Figure 2.4). A human body also emits "heat" radiation, with a maximum at $\lambda \approx 10$ μm. Thermal detectors for humans are, therefore, designed such that they are sensitive to radiation in the wavelength range 7–14 μm. NOAA's thermal scanner, with its interest in heat issuing from the Earth's surface, detects thermal IR radiation in the range 3.5– 12.5 μm. Object temperature is a kind of quantity often needed for studying a variety of environmental problems, as well as being useful for analysing the mineral composition of rocks and the evapotranspiration of vegetation.

thermal infrared

2.4   Interaction of atmosphere and EM radiation

Before the Sun's radiation reaches the Earth's surface, three RS-relevant interactions in the atmosphere have occurred: absorption, transmission, and scattering. The transmitted radiation is then either absorbed by the surface material or reflected. Before reaching a remote sensor, the reflected radiation is also subject to scattering and absorption in the atmosphere (Figure 2.6).
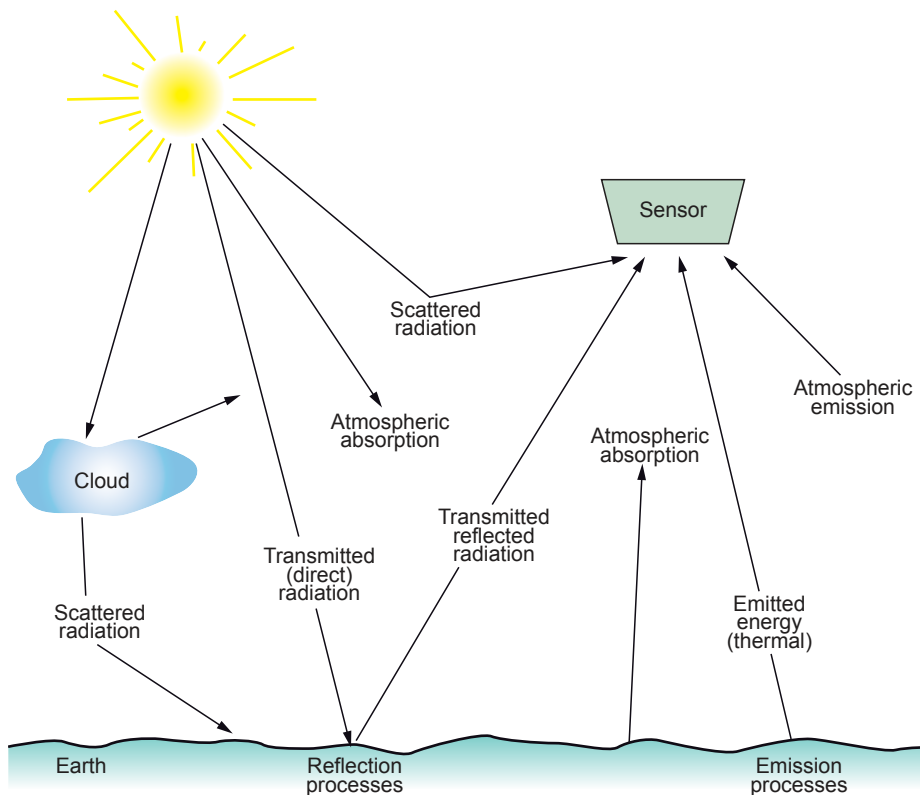
Figure 2.6
Interactions of EM radiation with the atmosphere and the Earth's surface.

2.4.1   Absorption and transmission

As it moves through the atmosphere, EM radiation is partly absorbed by various molecules. The most efficient absorbers of solar radiation in the atmosphere are ozone (O_3), water vapour (H_2O) and carbon dioxide (CO_2).
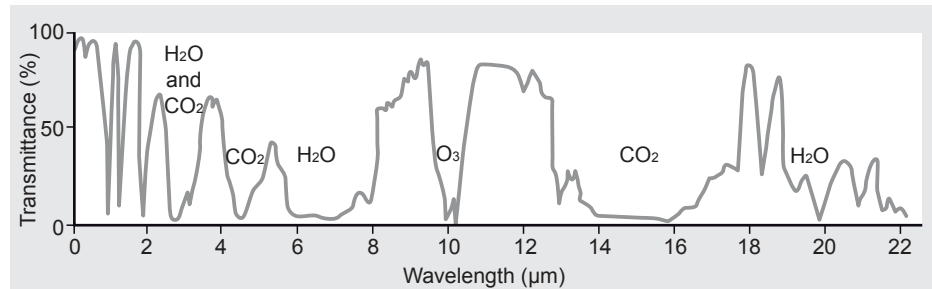
Figure 2.7
Atmospheric transmittance.

Figure 2.7 shows a schematic representation of atmospheric transmission in the wavelength range 0–22 µm. From this figure it can be seen that many of the wavelengths are not useful for remote sensing of the Earth's surface, simply because the corresponding radiation cannot penetrate the atmosphere. Only those wavelengths outside the main absorption ranges of atmospheric gases can be used for remote sensing. The useful ranges are referred to as *atmospheric transmission windows* and include:

atmospheric transmission

- the window from 0.4 to 2 µm. The radiation in this range (visible, NIR, SWIR) is mainly reflected radiation. Because this type of radiation follows the laws of optics, remote sensors operating in this range are often referred to as optical sensors.

- three windows in the TIR range, namely two narrow windows around 3 and 5 µm, and a third, relatively broad window extending from approximately 8 µm to 14 µm.

Because of the presence of atmospheric moisture, strong absorption occurs at longer wavelengths. There is hardly any transmission of radiation in the range from 22 µm to 1 mm. The more or less "transparent" range beyond 1 mm is the microwave range.

Solar radiation observed both with and without the influence of the Earth's atmosphere is shown in Figure 2.8. Solar radiation measured outside the atmosphere resembles black-body radiation at 6000 K. Measuring solar radiation at the Earth's surface shows that there the spectral distribution of the solar radiation is very ragged. The relative dips in this curve indicate the absorption by different gases in the atmosphere. We also see from Figure 2.8 that the total intensity in this range (i.e. the area under the curve) has decreased by the time the solar energy reaches the Earth's surface, after having passed through the atmosphere.

2.4.2 Atmospheric scattering

Atmospheric scattering occurs when particles or gaseous molecules present in the atmosphere cause EM radiation to be redirected from its original path. The amount of scattering depends on several factors, including the wavelength of the radiation in relation to the size of particles and gas molecules, the amount of particles and gases, and the distance the radiation travels through the atmosphere. On a clear day the colours are bright and crisp, and approximately 95% of the sunlight detected by our eyes, or a comparable remote sensor, is radiation reflected from objects; 5% is light scattered in the atmosphere. On a cloudy or hazy day, colours are faint and most of the radiation received by our eyes is scattered light. We may distinguish three types of scattering according to the size of particles in the atmosphere causing it. Each has a different relevance to RS.
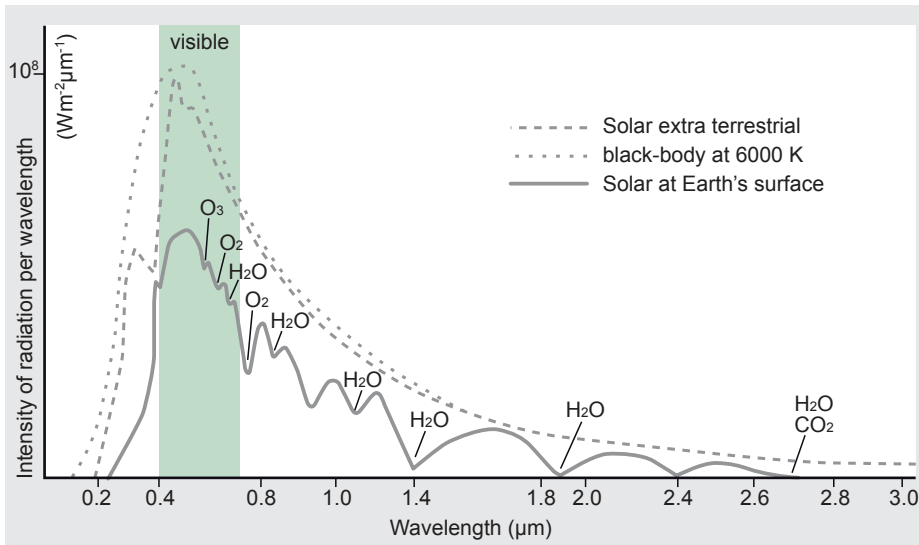
Rayleigh scattering dominates where electromagnetic radiation interacts with particles that are smaller than the wavelengths of light. Examples of such particles are tiny specks of dust and molecules of nitrogen (NO_2) and oxygen (O_2). Light of shorter wavelengths (e.g. blue) is scattered more than light of longer wavelengths (e.g. red); see Figure 2.9.

Rayleigh scattering
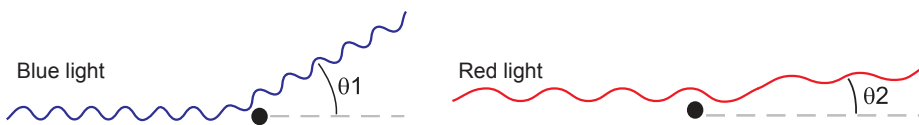
Figure 2.9
Rayleigh scattering is caused
by particles smaller than the
wavelengths of light and is
greater for small wavelengths.

In the absence of particles and scattering, the sky would appear black. During the day, solar radiation travels the shortest distance through the atmosphere; Rayleigh scattering causes a clear sky to be observed as blue. At sunrise and sunset, the sunlight travels a longer distance through the Earth's atmosphere before reaching the surface. All the radiation of shorter wavelengths is scattered after some distance and only the longer wavelengths reach the Earth's surface. As a result we do not see a blue but an orange or red sky (Figure 2.10).
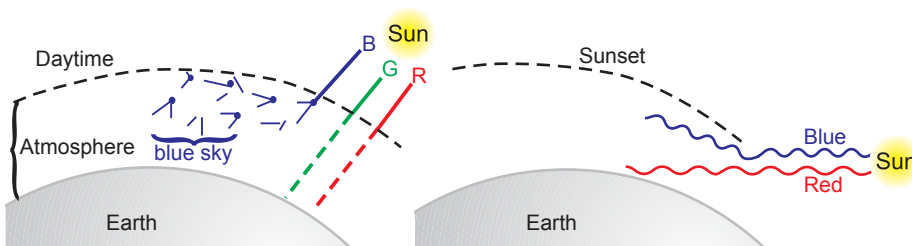
Figure 2.10
Rayleigh scattering causes
us to see a blue sky during
the day and a red sky at
sunset.

Rayleigh scattering disturbs RS in the visible spectral range from high altitudes. It causes a distortion of the spectral characteristics of the reflected light as compared to measurements taken on the ground: due to Rayleigh scattering, the shorter wavelengths are overestimated. This accounts for the blueness of colour photos taken from

high altitudes. In general, Rayleigh scattering diminishes the "crispness" of photos and thus reduces their interpretability. Similarly, Rayleigh scattering has a negative effect on digital classification using data from multispectral sensors.

Mie scattering occurs when the wavelength of EM radiation is similar in size to particles in the atmosphere. The most important cause of Mie scattering is the presence of aerosols: a mixture of gases, water vapour and dust. Mie scattering is generally restricted to the lower atmosphere, where larger particles are more abundant, and it dominates under overcast, cloudy conditions. Mie scattering influences the spectral range from the near-UV up to mid-IR range and has a greater effect on radiation of longer wavelengths than Rayleigh scattering.

Non-selective scattering occurs when particle sizes are much larger than the radiation wavelength. Typical particles responsible for this effect are water droplets and larger dust particles. Non-selective scattering is independent of the wavelength within the optical range. The most prominent example of non-selective scattering is that we see clouds as white bodies. A cloud consists of water droplets; since they scatter light of every wavelength equally, a cloud appears white. A remote sensor like our eye cannot "see through" clouds. Moreover, clouds have a further limiting effect on optical RS: clouds cast shadows (Figure 2.11).

Mie scattering

non-selective scattering
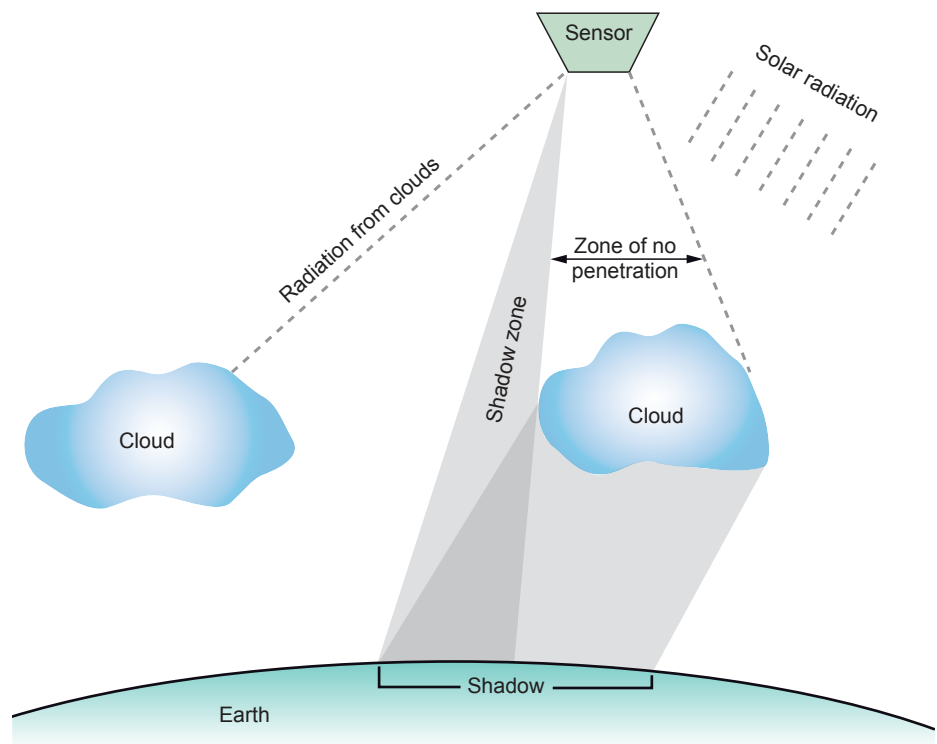
Figure 2.11
Direct and indirect effects of clouds for optical remote sensing.

2.5 Interactions of EM radiation with the Earth's surface

The EM radiation that reaches an object interacts with it. As a result of this interaction, EM radiation is absorbed, transmitted or reflected by the object. The energy conservation law, applied to interaction of EM radiation with the object, states that *all* incident EM radiation (I) is absorbed (A), reflected (R), or transmitted (T):

$$A(\lambda) + R(\lambda) + T(\lambda) = I(\lambda) \tag{2.7}$$

It is important to note that Equation 2.7 applies for each wavelength. Dividing both sides of Equation 2.7 by I we get.

$$\frac{A(\lambda)}{I(\lambda)} + \frac{R(\lambda)}{I(\lambda)} + \frac{T(\lambda)}{I(\lambda)} = \alpha(\lambda) + \rho(\lambda) + \tau(\lambda) = 1 \tag{2.8}$$

where $\alpha(\lambda)$ is absorptance, $\rho(\lambda)$ is reflectance and $\tau(\lambda)$ is transmittance of the object, all depend on wavelength λ and range from 0 to 1. For opaque objects $\tau(\lambda) = 0$ and Equation 2.8 reduces to

$$\alpha(\lambda) + \rho(\lambda) = 1 \tag{2.9}$$

Absorption of EM radiation leads to an increase in the object's temperature, while emission of EM radiation leads to a decrease in the object's temperature. The amount of emitted EM radiation is determined by the object's temperature (see Planck's law) and emissivity $\epsilon(\lambda)$. In equilibrium the total amounts of absorbed and emitted radiation at all wavelength are equal and the object's temperature is constant.

equilibrium

Kirchhoff's law of thermal radiation states that in equilibrium absorptance and emissivity at each wavelength are equal:

Kirchhoff's law

$$\alpha(\lambda) = \epsilon(\lambda) \tag{2.10}$$

The reflectance, transmittance and absorptance will vary with wavelength and type of target material. Here and further in the book we define a *target* as an object on the Earth surface that is being detected or sensed. Also the surface of target influences interaction of EM radiation and the target. Two types of reflection that represent the two extremes of the way in which radiation is reflected by a target are "specular reflection" and "diffuse reflection" (Figure 2.12). In the real world, usually a combination of both types is found.

reflection
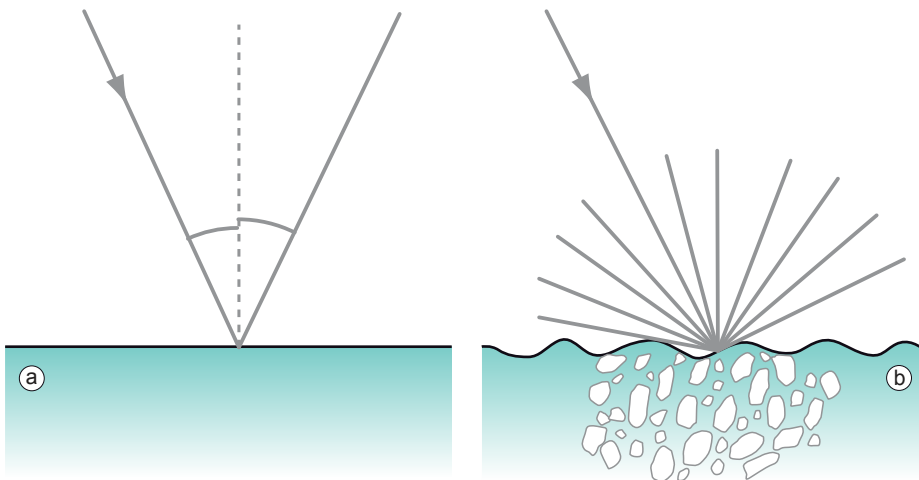
Figure 2.12
Schematic diagrams showing (a) specular and (b) diffuse reflection.

- *Specular reflection*, or mirror-like reflection, typically occurs when a surface is smooth and (almost) all of the radiation is directed away from the surface in a

single direction. Specular reflection can occur, for example, for a water surface or a glasshouse roof. It results in a very bright spot (also called "hot spot") in the sensed image.

- *Diffuse reflection* occurs in situations where the surface is rough and the radiation is reflected almost uniformly in all directions.

Whether a particular target reflects specularly, diffusely, or both, depends on the surface roughness relative to the wavelength of the incident radiation.

2.5.1 Spectral reflectance curves

We can establish for each type of material of interest a *reflectance curve*. Such a curve shows the portion of the incident radiation ρ that is reflected as a function of wavelength λ (expressed as percentage; see Figure 2.13). Remote sensors are sensitive to ranges, albeit narrow, of wavelengths, not just to one particular λ, for example the "spectral band" from $\lambda = 0.4$ μm to $\lambda = 0.5$ μm. The spectral reflectance curve can be used to estimate the overall reflectance in such bands by calculating the mean of reflectance measurements in the respective ranges. Reflectance measurements can be carried out in a laboratory or in the field, in the latter case using a field spectrometer. Reflectance curves are typically collected for the optical part of the electromagnetic spectrum and large efforts are made to store collections of typical curves in "spectral libraries". The reflectance characteristics of some common land cover types are discussed in the following subsections.

reflectance measurements

Vegetation

The reflectance characteristics of vegetation depend on the properties of the leaves, including the orientation and structure of the leaf canopy. The amount of radiation reflected for a particular wavelength depends on leaf pigmentation, thickness and composition (cell structure), and on the amount of water in the leaf tissue. Figure 2.13 shows an ideal reflectance curve of healthy vegetation. In the visible portion of the spectrum, the reflection of the blue and red components of incident light is comparatively low, because these portions are absorbed by the plant (mainly by chlorophyll) for photosynthesis; the vegetation reflects relatively more green light. The reflectance in the NIR range is highest, but the amount depends on leaf development and cell structure. In the SWIR range, reflectance is mainly determined by the free water in the leaf tissue; more free water results in less reflectance. Wavelengths around 1.45 μm and 1.95 μm are, therefore, called water absorption bands. The plant may change colour when its leaves dry out, for instance at harvest time for a crop (e.g. to yellow). At this stage there is no photosynthesis, which causes reflectance in the red portion of the spectrum to become higher. Also, the leaves will dry out, resulting in a higher reflectance of SWIR radiation, whereas reflectance in the NIR range may decrease. As a result, optical remote sensing can provide information about the type of plant and also about its health.

Bare soil

Reflectance from bare soil depends on so many factors that it is difficult to give one typical soil reflectance curve. The main factors influencing reflectance are soil colour, moisture content, the presence of carbonates, and iron oxide content. Figure 2.14 gives the reflectance curves for the five main types of soil occurring in the U.S.A. Note the typical shapes of most of the curves, which are convex shape in the range 0.5–1.3 μm and dip at 1.45 μm and 1.95 μm. These dips correspond to water absorption bands and are caused by the presence of soil moisture. Iron-dominated soil (e) has quite a different reflectance curve since iron absorption dominates at longer wavelengths.
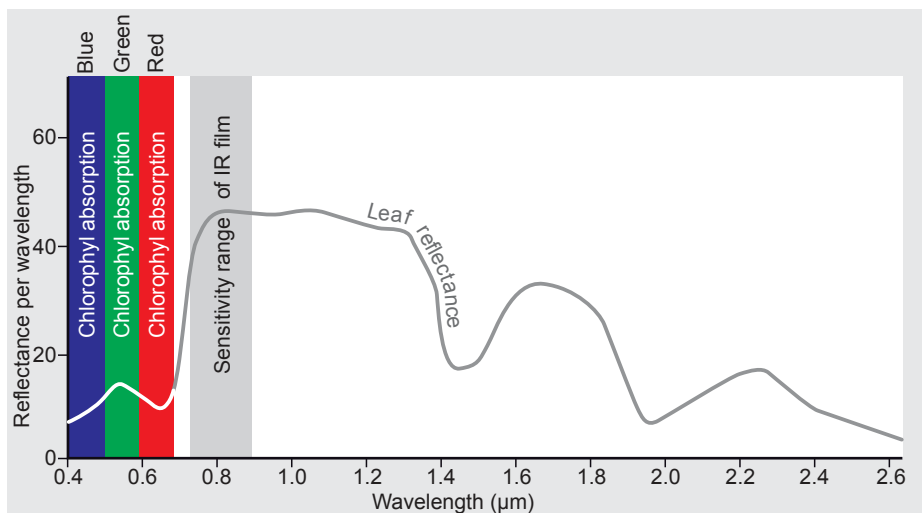
Figure 2.13
An idealized spectral reflectance curve of healthy vegetation.
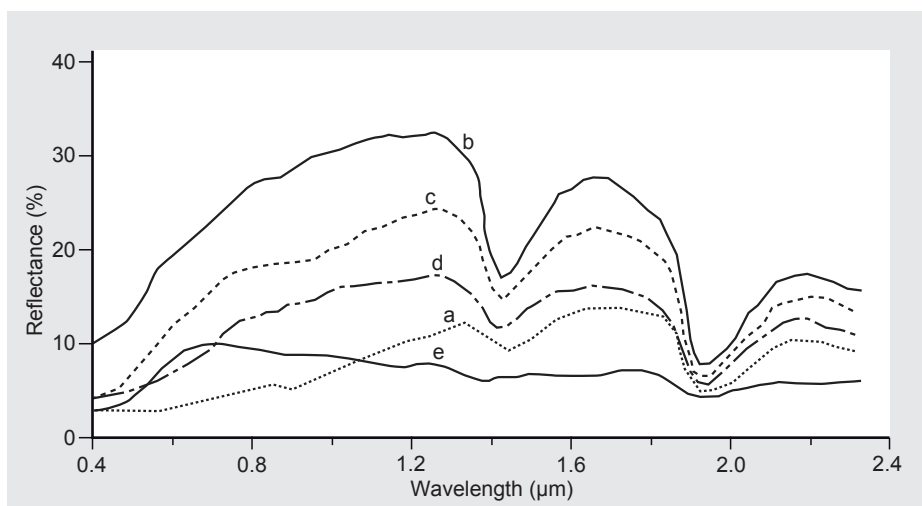
Figure 2.14
Spectral reflectance of five mineral soils: (a) organic dominated, (b) minimally altered, (c) iron altered, (d) organic affected and (e) iron dominated (from [71]).
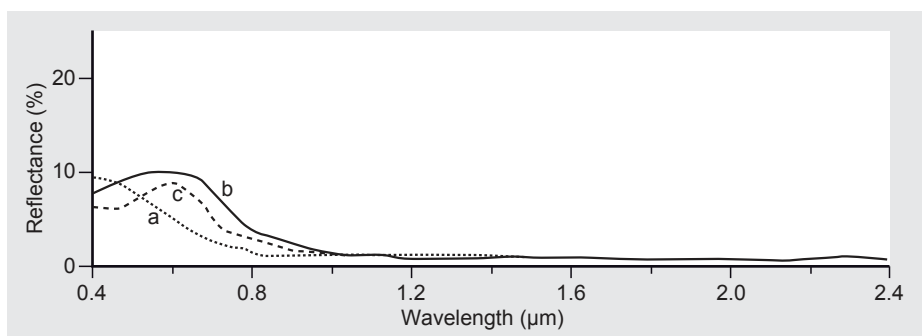
Figure 2.15
Typical effects of chlorophyll and sediments on water reflectance: (a) ocean water, (b) turbid water, (c) water with chlorophyll (from [71]).

Water

Compared to vegetation and soils, water has a lower reflectance. Vegetation may reflect up to 50% and soils up to 30–40%, while water reflects at most 10% of the incident

radiation. Water reflects EM radiation in the visible range and a little in the NIR range. Beyond 1.2 µm, all radiation is absorbed. Spectral reflection curves for water of different compositions are given in Figure 2.15. Turbid (silt loaded) water has the highest reflectance. Water containing plants or algae has a pronounced reflectance peak for green light because of the chlorophyll present.

2.6 Sensing of EM radiation

The review of properties of EM radiation shows that different forms of radiation can provide us with different information about terrain-surface features and that different applications of Earth Observation are likely to benefit from sensing in different ranges of the EM spectrum. A geoinformatics engineer who wants to discriminate objects for topographic mapping will prefer to use an optical sensor operating in the visible range. An environmentalist who needs to monitor heat losses of a nuclear power plant will use a sensor that detects thermal emission. A geologist interested in surface roughness, because it indicates to him rock type, will rely on microwave sensing. Different demands combined with different technical solutions have resulted in a multitude of sensors. In this section we will classify various remote sensors and discuss their common features. Peculiarities will then be treated later in appropriate sections.

2.6.1 Sensing properties

A *remote sensor* is a device that detects EM radiation, quantifies it and, usually, records it in an analogue or digital form. A remote sensor may also transmit recorded data (to a receiving station on the ground). Many sensors used in Earth Observation detect reflected solar radiation. Others detect the radiation emitted by the Earth itself. There are, however, some obstacles to be overcome. The Sun does not always shine brightly and there are regions on the globe almost permanently under cloud cover. There are also regions that have seasons with very low Sun elevation, so that objects cast long shadows over long periods. Furthermore, at night there are only emissions—
obstacles to sensing   and perhaps moonlight. Sensors detecting reflected solar radiation are useless at night and face problems when dealing with unfavourable seasonal and weather conditions. Sensors detecting emitted terrestrial radiation do not directly depend on the Sun as a source of illumination; they can be operated any time. The Earth's emissions, we have learned, occurs only at longer wavelengths because of the relatively low surface temperature and because long EM waves do not hold much energy, which makes them more difficult to sense.

Luckily we do not have to rely only on solar and terrestrial radiation. We can build instruments that emit EM radiation and then detect the radiation returning from the target object or surface. Such instruments are called *active sensors*, as opposed to passive ones, which measure reflected solar or terrestrial radiation (Figure 2.16). An example of an active sensor is a laser rangefinder, a device that can be bought for a few euros in any DIY store. Another very common active sensor is a camera with a flash
active versus passive RS   unit (which will operate below certain levels of light). The same camera without the flash unit is a passive sensor. The main advantages of active sensors are that they can be operated day and night and have a controlled illuminating signal. They are often designed to work in an EM spectrum range that is less affected by the atmosphere and weather conditions. Laser and radar instruments are the most prominent active sensors for GDA.

Most remote sensors measure either the intensity or the phase of EM radiation. Some— like a simple laser rangefinder—only measure the elapsed time between sending a radiation signal and receiving it back. Radar sensors may measure both intensity and
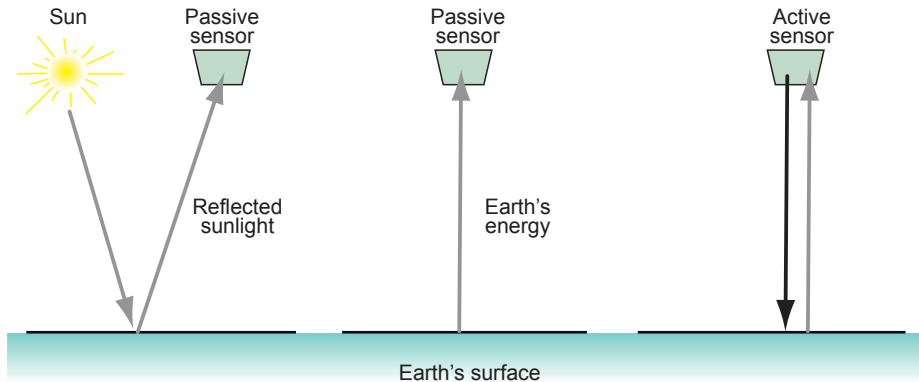
phase. Phase measuring sensors are used for precise ranging (distance measurement), e.g. by GPS "phase receivers" or continuous-wave laser scanners. The intensity of radiation can be measured from the photon energy striking the sensor's radiation-sensitive surface.

intensity or phase

By considering the following equation, you can relate the intensity measure of reflected radiation to Figure 2.6 and link the Figures 2.13 to 2.17. When sensing reflected light, radiance at the sensor is equal to the radiance at the Earth's surface attenuated by atmospheric absorption, plus the radiance of scattered light:

measuring radiance

$$L = \frac{\rho E \tau}{\pi} + \text{sky radiance} \qquad (2.11)$$

where L is the total radiance at the sensor, E is the irradiance (the intensity of the incident solar radiation, attenuated by the atmosphere) at the Earth's surface, ρ is the terrain reflectance, and τ is the atmospheric transmittance. The radiance at the Earth's surface depends on the irradiance and the terrain surface reflectance. The irradiance, in turn, stems from direct sunlight and diffuse light, the latter caused by atmospheric scattering, particularly on a hazy days (see Figure 2.17). This indicates why you should study radiometric correction (Subsection 5.1.3 and Subsection 5.2.2), to enable you to make better inferences about surface features.

The radiance is observed for a *spectral band*, not for a single wavelength. A *spectral band* or wavelength band is an interval of the EM spectrum in which the average radiance is measured. Sensors such as a panchromatic camera, a radar sensor and a laser scanner only measure in one specific band, while a multispectral scanner or a digital camera measures in several spectral bands at the same time. Multispectral sensors have several *channels*, one for each spectral band. Figure 2.18 shows spectral reflectance curves, together with the spectral bands, of some popular satellite-based sensors. Sensing in several spectral bands simultaneously allows us to relate properties that show up well in specific spectral bands. For example, reflection characteristics in the spectral band 2 to 2.4 µm (as recorded by Landsat-5 TM channel 7) tell us something about the mineral composition of soil. The combined reflection characteristics in the red and NIR bands (from Landsat-5 TM channels 3 and 4) can tell us something about biomass and plant health.

spectral band

Landsat MSS (MultiSpectral Scanner), the first civil space-borne Earth Observation sensor, had sensing elements (detectors) for three rather broad spectral bands in the visible range of the spectrum, each with a width of 100 nm, and one broader band in the NIR range. A hyperspectral scanner uses detectors for many more, but narrower, bands, which may be as narrow as 20 nm, or even less. We say a hyperspectral sensor
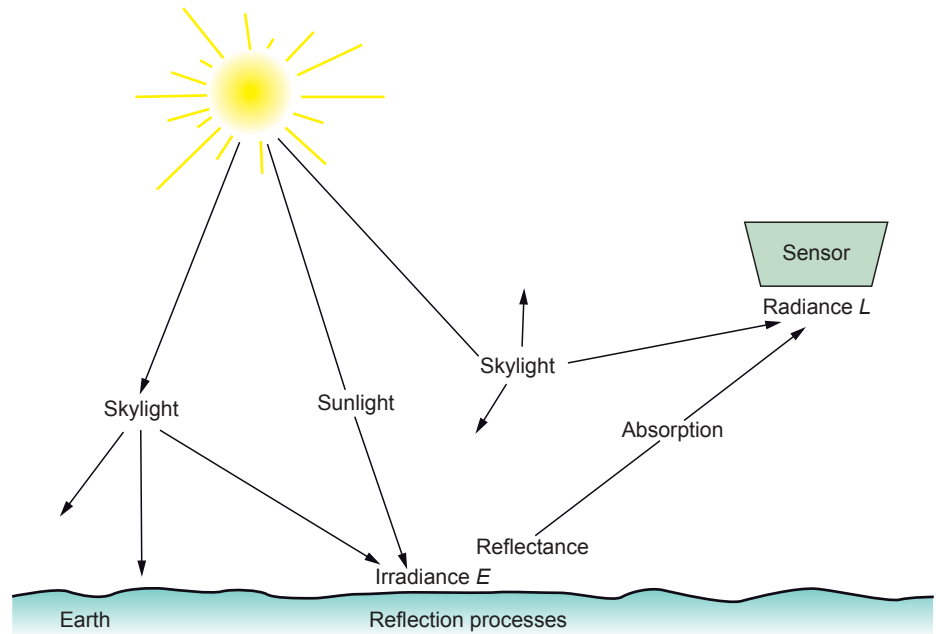
Figure 2.17
Radiance at the sensor,
adapted from [67].
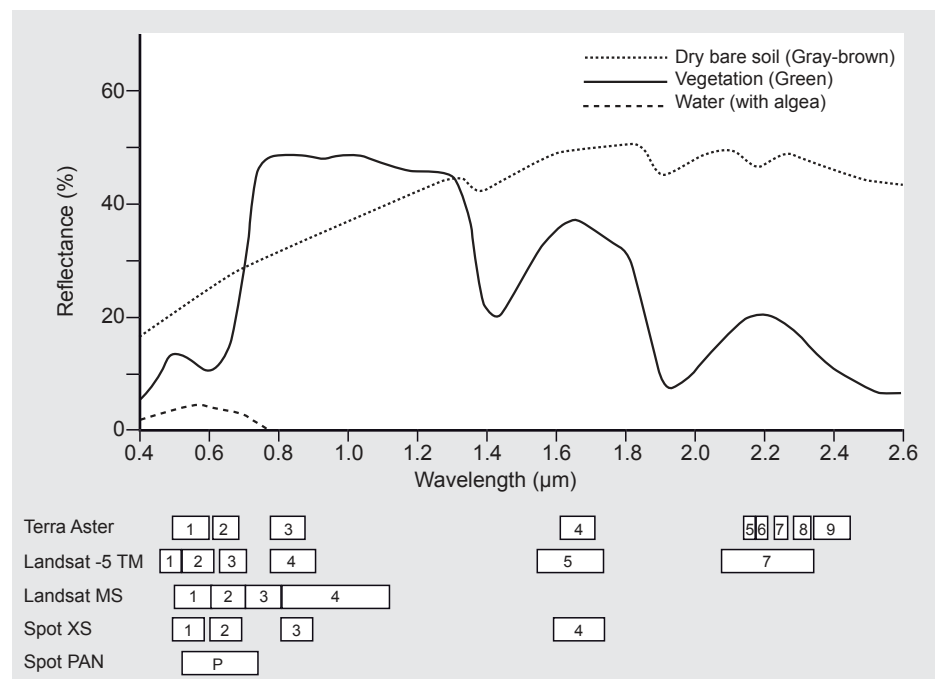
Figure 2.18
Spectral reflectance curves
and spectral bands of some
multispectral sensors.

spectral resolution

has a higher 'spectral resolution' than a multispectral one. A laser instrument can emit
(and detect) almost monochrome radiation, with a wavelength band no wider than
10 nm. A camera loaded with panchromatic film or a space-borne electronic sensor
with a panchromatic channel (such as SPOT PAN or WorldView-1) records the inten-
sity of radiation of a broad spectral band covering the entire visible range of the EM
spectrum. Panchromatic—which stands for "across all colours"—recording is compa-

rable with the function of the 120 million rods of a human eye. They are brightness sensors and cannot sense colour.

In a camera loaded with panchromatic film (*black & white film*), the silver halide crystals of the light-sensitive emulsion detect radiation. The silver halide grains turn to silver metal when exposed to light, the more so the higher the intensity of the incident light. Each light ray from an object/scene triggers a chemical reaction of some particular grain. This way, variations in radiance within a scene are detected and an image of the scene is created at the time of exposure. The record obtained is only a latent image; the film has to be developed to turn it into a photograph.

photographic detector

Digital cameras and multispectral scanners are examples of sensors that use electronic detectors instead of photographic ones. An electronic detector (CCD, CMOS, photodiode, solid state detector, etc.) is made of semiconductor material. The detector accumulates a charge by converting the photons incident upon its surface to electrons. (It was Einstein who won the Nobel prize for discovering and explaining that there is an emission of electrons when a negatively charged plate of light-sensitive (semiconductor) material is subject to a stream of photons.) The electrons can then be made to flow as a current from the plate. So the charge can be converted to a voltage (electrical signal). The charge collected is proportional to the radiance at the detector (the amount of radiation "deposited" in the detector). In a process called A/D conversion, the electrical signal is *sampled* and quantified. The output is a digital number (DN), which is recorded. A DN is an integer within a fixed range. Older remote sensors used 8 bits for recording, which allows a differentiation of radiance into $2^8 = 256$ levels (i.e. DNs in the range 0 to 255). The recently launched (in 2007) WorldView-1 sensor records with a *radiometric resolution* of 11 bits $(2^{11} = 2048)$. ASTER records the visible spectral band using 8 bits and the thermal infrared band using 12 bits. A higher radiometric resolution requires more storage capacity but has the advantage of offering data with greater information content (see Figure 2.19).

AD conversion

radiometric resolution
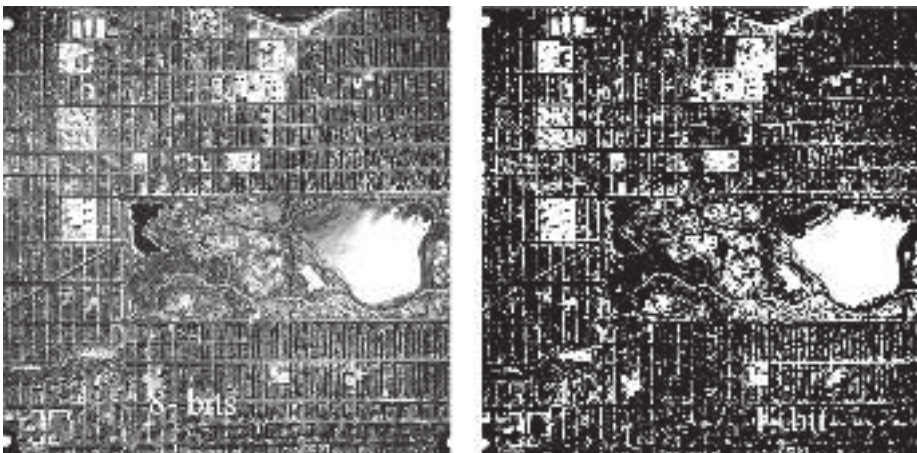
Figure 2.19
8-bit versus 11-bit radiometric resolution.

A digital panchromatic camera has an array of detectors instead of silver halide crystals suspended in gelatine on a polyester base of photographic film. Each detector (e.g. a CCD, which stands for charge-coupled device) is very small, in the order of 9 μm × 9 μm. Space-borne cameras use larger detectors than aerial cameras to ensure that enough photons are collected despite the great distances at which they operate from the Earth. At the moment of exposure, each detector yields one DN, so in total we obtain a data set that represents an image similar to the one created by "exciting" photographic material in a film camera.

imaging,
spatial resolution

When arranging the DNs in a two-dimensional array, we can readily visualize them as grey values. We refer to the obtained "image" as a *digital image* and to a sensor producing digital images as an *imaging sensor*. The array of DNs represents an image in terms of discrete picture elements, called *pixels*. The value of a pixel—its DN—corresponds to the radiance of the light reflected from the small ground area viewed by the relevant detector. The smaller the detector, the smaller will be the area on the ground that corresponds to one pixel. The size of the "ground resolution cell" is often referred to as "pixel size on the ground". Early digital cameras for consumers had 2×10^6 CCDs per spectral band (named 2 megapixel cameras); today we can get for the same price a 10 megapixel camera. The latter has much smaller CCDs so that they can fit on the same board, with the consequence that an image can reveal much more detail; we would say the *spatial resolution* of the image is higher.

A digital camera for the consumer market does not record intensity values for a single (panchromatic) spectral band, but for three bands simultaneously, namely for red, green, and blue light, in order to obtain colour images. This is comparable with our eyes: we have three types of cones, one for each primary colour. The data set obtained for one shot taken with the camera (the *image file*) therefore contains three separate digital images (Figure 2.20). Multispectral sensors record in as many as 14 bands simultaneously (e.g. ASTER). For convenience, a single digital image is then often referred to as "band" and the total image file as a *multi-band image*.
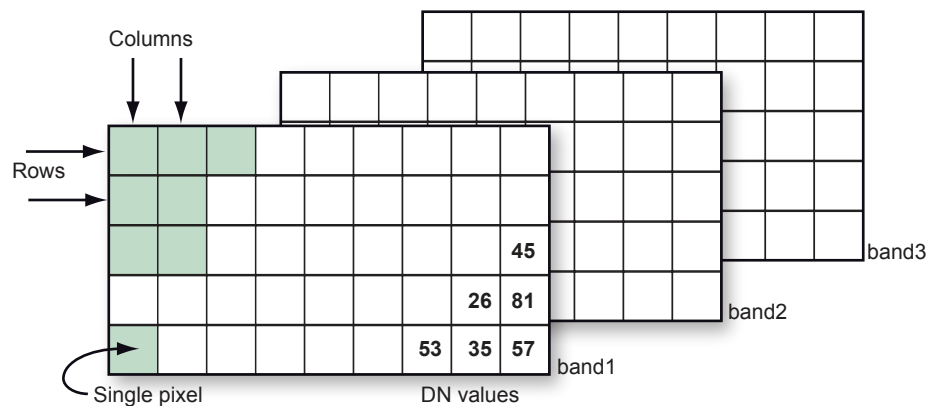
Figure 2.20
An image file comprises a digital image for each of the spectral bands of the sensor. The DN values for each band are stored in a row-column arrangement.

storage media

Various storage media are used for recording the huge amount of data produced by electronic detectors: solid state media (such as memory cards as used in consumer cameras), magnetic media (disk or tape) and optical discs (some video cameras); satellites usually have several recorders on board.

Light sensor systems often transmit data to ground receiving stations at night. Data can also be transmitted directly to a receiving station using satellite communication technology. Airborne sensors often use the hard disk of a laptop computer as a recording device. The huge amounts of data collected demand efficient data management systems. This issue will be examined in Section 8.4.

2.6.2 Classification of sensors

Remote sensors can be classified and labelled in different ways. According to whatever our prime interest in Earth Observation may be—geometric properties, spectral differences, or an intensity distribution of an object or scene—we can distinguish three salient types of sensors: altimeters, spectrometers, and radiometers.

Laser and radar altimeters are non-imaging sensors that provide information about

the elevation of water and land surfaces.

Thermal sensors, such as the channels 3 to 5 of NOAA's AVHRR or the channels 10 to 14 of Terra's ASTER, are called (imaging) radiometers. Radiometers measure radiance and typically sense in one broad spectral band or in only a few bands, but with high radiometric resolution. Panchromatic cameras and passive microwave sensors are other examples of radiometers. The spatial resolution depends on the wavelength band of the sensor. Panchromatic radiometers can have a very high spatial resolution, whereas microwave radiometers have a low spatial resolution because of the low levels of energy inherent in this spectral range. Scatterometers are non-imaging radiometers. Radiometers are used for a wide range of applications: for example, detecting forest/bush/coal fires; determining soil moisture and plant response; monitoring ecosystem dynamics; and analysing energy balance across land and sea surfaces.

Spectrometers measure radiance in many (usually about 100 or 200) narrow, contiguous spectral bands and therefore have a high spectral resolution. Their spatial resolution is moderate to low. The prime use of imaging spectrometers is to identify surface materials—from the mineral composition of soils, to concentrations of suspended matter in surface water and chlorophyll content. There are also androgynous sensors: spectro-radiometers, imaging laser scanners, and Fourier spectrometers, for example.

We can also group the multitude of remote sensors used for GDA according to the spectral domains in which they operate (Figure 2.21). The following list gives a short description of each group and refers to the section in which they are treated in more detail.
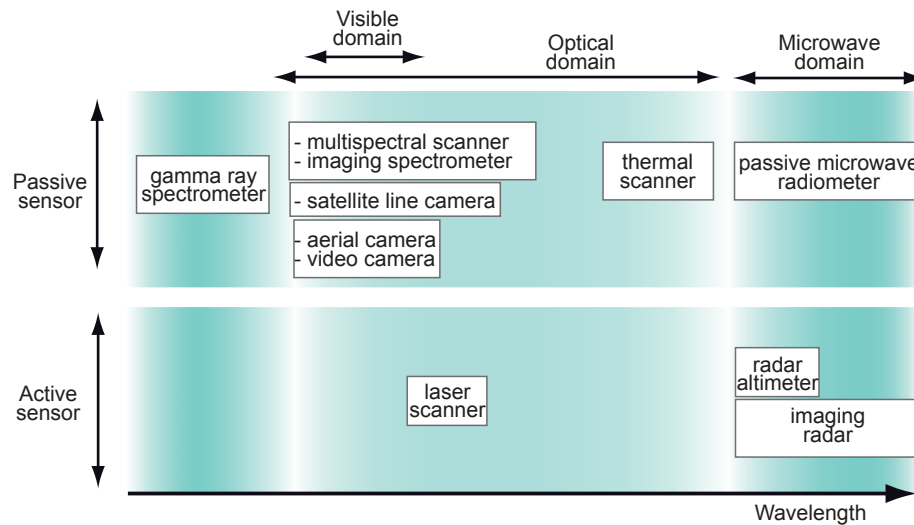
Figure 2.21
Overview of the sensors that are described in this book.

- gamma ray spectrometers are mainly used in mineral exploration.

- aerial film cameras have been the remote sensing workhorse for decades. Today, they are used primarily for large-scale topographic mapping, cadastral mapping, and orthophoto production for urban planning, to mention a few examples; they are discussed in Section 4.6.

- digital aerial cameras are not conquering the market as quickly as digital cameras did on the consumer market. These cameras use CCD arrays instead of film; they are treated together with optical scanners in Section 4.1. Line cameras operated from satellites have very similar properties.

video cameras

- digital video cameras are not only used to record movies. They are also used in aerial Earth Observation to provide low cost (and low resolution) images for mainly qualitative purposes, for instance to provide visual information about an area covered by "blind" airborne laser scanner data. Handling images from video cameras is similar to dealing with images from digital "still" cameras; this is not explicitly discussed any further in this book.

multispectral scanners

- multispectral scanners are mostly operated from satellites and other space vehicles. The essential difference between multispectral scanners and satellite line cameras is the imaging/optical system employed: multispectral scanners use a moving mirror to "scan" a line (i.e. a narrow strip on the ground) and a single detector instead of recording intensity values of an entire line at one instant by an array of detectors as for line cameras. Multispectral scanners are treated in Section 4.1. Figure 2.22 shows an image obtained by combining the images of Landsat TM channels 4, 5 and 7, which are displayed in red, green and blue, respectively. Section 5.1 explains how to produce such a "false colour" image.
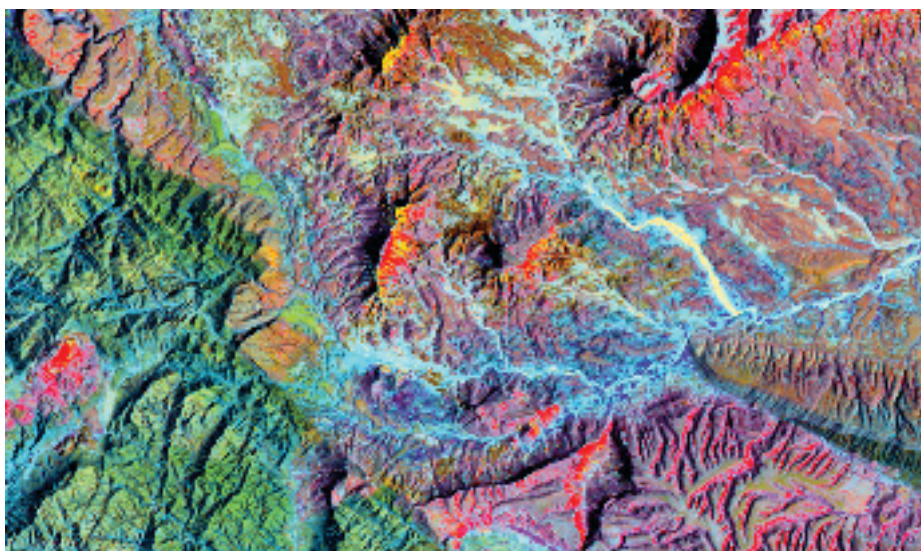
Figure 2.22
Landsat-5 TM false colour composite of an area of 30 km × 17 km.

imaging spectrometers

- hyperspectral scanners are imaging spectrometers with a scanning mirror; they are treated in detail in Section 4.3.

thermal scanners

- thermal scanners are placed here in the optical domain purely for the sake of convenience. They exist as special instruments and as a component of multispectral radiometers; they are included in Section 4.2. Thermal scanners provide us with data that can be directly related to object temperature. Figure 2.23 is an example of a thermal image acquired by an airborne thermal scanner at night.

microwave radiometers

- passive microwave radiometers detect emitted radiation of the Earth's surface in the 10 to 1000 mm wavelength range. These radiometers are mainly used in mineral exploration, for monitoring soil-moisture changes, and for snow and ice detection. Microwave radiometers are not discussed further in this book.

- laser scanners are the scanning variant of laser rangefinders and altimeters (as on ICESat). Laser scanners measure the distance from the laser instrument to many points of the target in "no time" (e.g. 150,000 points in one second). Laser
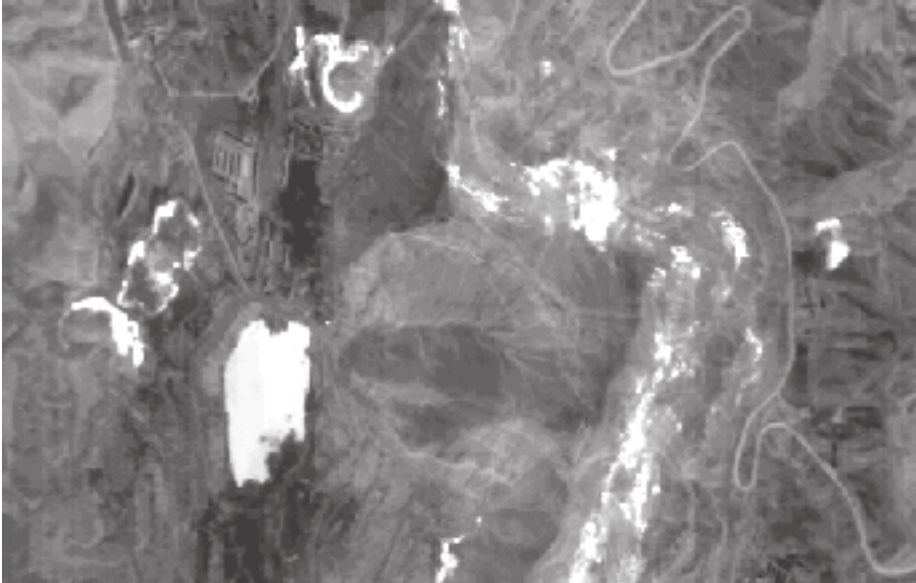
Figure 2.23
'Thermal image' at night of a coal mining area affected by underground coal fires. Darker tones represent colder surfaces, while lighter tones represent warmer areas. Most of the warm spots are due to coal fires, except for the large white patch, which is a lake; at that time of the night, apparently the temperature of the water was higher than the temperature of the land. On the ground, the area depicted is approximately 4 km across.

ranging is often referred to as LIDAR (LIght Detection And Ranging). The prime application of airborne laser scanning (ALS) is for creating high resolution digital surface models and digital terrain models (see Section 5.3). We can also create a digital terrain model (DTM) from photographs and similar panchromatic images. However, because of the properties of laser radiation, ALS has important advantages in areas of dense vegetation and for sandy deserts and coastal areas. Surface modelling is of interest for many applications, such as, for example, biomass estimation of forests, volume calculations for open-pit mining (see Figure 2.24), flood plain mapping, and 3D modelling of cities. Laser scanning is dealt with in more detail in Section 4.5.

Figure 2.24
Pictorial representation of a digital surface model of the Sint Pietersberg open-pit mine in the Netherlands. The size of the pit is roughly 2 km × 1 km. The terraced rim of the pit is clearly visible. The black strip near the bottom of the image is the River Meuse. Courtesy Survey Department, Rijkswaterstaat.

• imaging radar (RAdio Detection And Ranging) operates in the spectral range

10–1000 mm. Radar instruments are active sensors and because of the range of wavelengths used they can provide data day and night, under all weather conditions. Radar waves can penetrate clouds; only heavy rainfall affects imaging to some degree. One of its applications is, therefore, the mapping of areas that are subject to permanent cloud cover. Figure 2.25 shows an example of a SAR (Synthetic Aperture Radar) image from the ERS-1 satellite. Radar data from the air or space can also be used to create surface models. Radar imaging has a peculiar geometry and processing raw radar data is not simple. Radar is treated in Section 4.4.

Figure 2.25
An ERS-1 SAR image of the Mahakam Delta, Kalimantan. The image shows different types of land cover. The river is black. The darker patch of land on the the left is inland tropical rainforest. The rest is a mixed forest of Nipa palm and mangrove on the delta. The right half of the image shows light patches, where the forest has been partly cleared. The image covers an area on the ground of 30 km x 15 km.
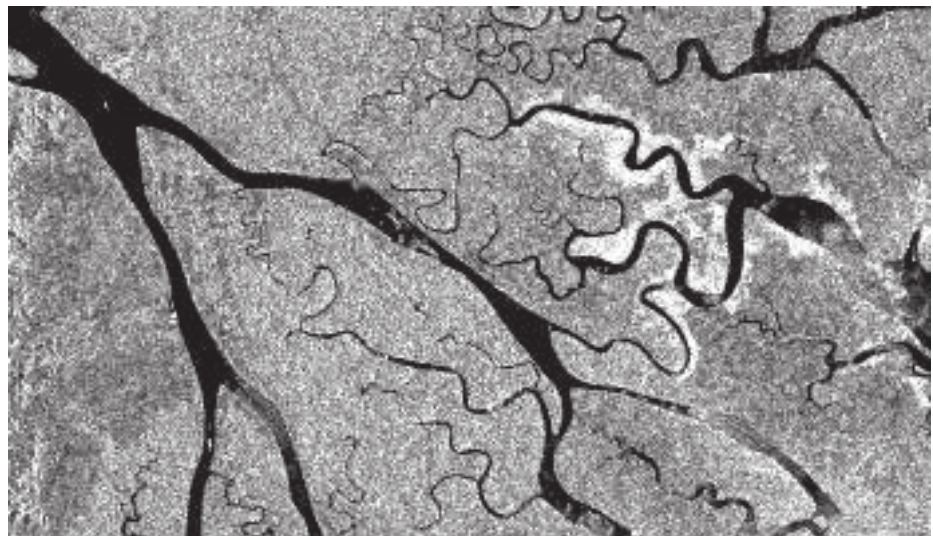
radar altimeters

- radar altimeters are used to measure elevation profiles of the Earth's surface that is parallel to the receiving satellite's orbit. Radar altimeters operate in the 10–60 mm range and allow us to calculate elevation with an accuracy of 20–50 mm. Radar altimeters are useful for measuring relatively smooth surfaces.

sonar

- for the sake of completeness, sonar, another active sensor, is included here. Sonar, which stands for SOund NAvigation Ranging, is used, for example, for mapping river beds and sea floors, and for detecting obstacles underwater. Sonar works by emitting a small burst of sound from a ship. The sound is reflected off the bottom of the body of water. The time taken for the reflected pulse to be received corresponds to the depth of the water. More advanced systems also record the intensity of the return signal, thus giving information about the material on the sea floor. In its simplest form, sonar "looks" vertically and is operated very much like a radar altimeter. The body of water will be traversed in paths resembling a grid; not every point below the water surface will be monitored. The distance between data points depends on the ship's speed, the frequency of the measurements, and the distance between the adjacent paths.

One of the most accurate systems for imaging large areas of the ocean floor is side-scan sonar. It is an imaging system that works in a way that is somewhat similar to side-looking airborne radar (see Section 4.4). The images produced by side-scan sonar systems are highly accurate and can be used to delineate even very small (< 1 cm) objects. From sonar data, we can produce contour maps of sea floors and other water bodies, which can be used, for example, for navigation and water-discharge studies.

Chapter 3

Spatial referencing and satellite-based positioning

Richard Knippers
Klaus Tempfli

Introduction

In the early days of geoinformation science, spatially referenced data usually originated within national boundaries, i.e. these data were derived from printed maps published by national mapping organizations. Nowadays, users of geoinformation are combining spatial data from a given country with global spatial data sets, reconciling spatial data from published maps with coordinates established by satellite positioning techniques, and integrating their spatial data with that from neighbouring countries.

To perform these kinds of tasks successfully, we need to understand basic spatial referencing concepts. Section 3.1 discusses the relevance and actual use of reference surfaces, coordinate systems and coordinate transformations. We will explain the principles of spatial referencing as applied to mapping, the traditional application of geoinformation science. These principles are generally applicable to all types of geospatial data.

Section 3.2 discusses satellite-based systems of spatial positioning. The development of these global positioning systems has made it possible to unambiguously determine any position in space. This and related developments have laid the foundations for the integration of all spatial data within a single, global 3D spatial-reference system, which we may expect to emerge in the near future.

3.1 Spatial referencing

One of the defining features of geoinformation science is its ability to combine spatially referenced data. A frequently occurring issue is the need to combine spatial data from different sources that use different spatial reference systems. This section provides

a broad background of relevant concepts relating to the nature of spatial reference systems and the translation of data from one spatial referencing system into another.

3.1.1 Reference surfaces

The surface of the Earth is far from uniform. Its oceans can be treated as reasonably uniform, but the surface or topography of its land masses exhibits large vertical variations between mountains and valleys. These variations make it impossible to approximate the shape of the Earth with any reasonably simple mathematical model. Consequently, two main reference surfaces have been established to approximate the shape of the Earth : one is called the *Geoid*, the other the *ellipsoid*; see Figure 3.1.

geoid and ellipsoid

Figure 3.1
The Earth's surface and two reference surfaces used to approximate it: the Geoid; and a reference ellipsoid. The Geoid separation (N) is the deviation between the Geoid and the reference ellipsoid.
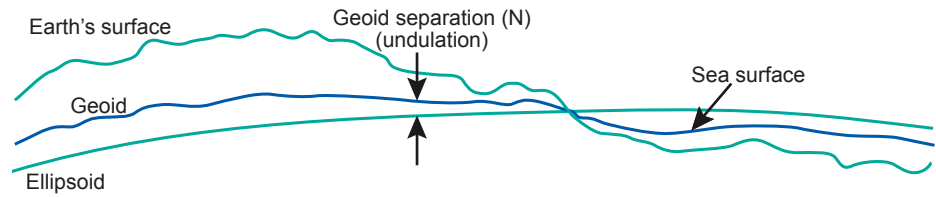
The Geoid and the vertical datum

We can simplify matters by imagining that the entire Earth's surface is covered by water. If we ignore effects of tides and currents on this *global ocean*, the resultant water surface is affected only by gravity. This has an effect on the shape of this surface because the direction of gravity—more commonly known as the plumb line—is dependent on the distribution of mass inside the Earth. Owing to irregularities or mass anomalies in this distribution, the surface of the *global ocean* would be undulating. The resulting surface is called the Geoid (Figure 3.2). A plumb line through any surface point would always be perpendicular to the surface.

plumb line
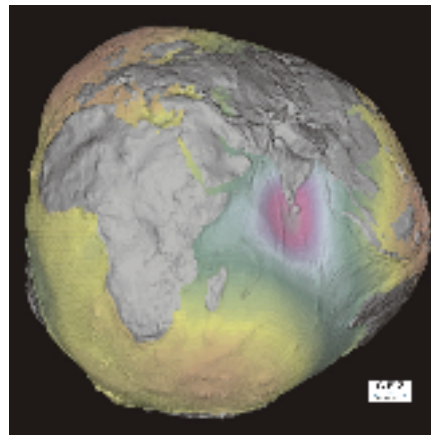
Figure 3.2
The Geoid, exaggerated to illustrate the complexity of its surface. Image: GFZ German Research Centre for Geosciences.

The Geoid is used to describe *heights*. In order to establish the Geoid as a reference for heights, the ocean's water level is registered at coastal locations over several years using tide gauges (mareographs). Averaging the registrations largely eliminates variations in sea level over time. The resultant water level represents an approximation to the Geoid and is termed mean sea level.

mean sea level

For the Netherlands and Germany, local mean sea level is related to the Amsterdam Tide Gauge (zero height). We can determine the height of a point in Enschede with

respect to the Amsterdam Tide Gauge using a technique known as geodetic levelling (Figure 3.3). The result of this process will be the height of the point in Enschede above local mean sea level. Height determined with respect to a tide gauge station is known as *orthometric height* (height H above the Geoid) .

Several definitions of local mean sea levels (also called local vertical datums) appear throughout the world. They are parallel to the Geoid but offset by up to a couple of metres to allow for local phenomena such as ocean currents, tides, coastal winds, water temperature and salinity at the location of the tide gauge. Care must be taken when using heights from another local vertical datum . This might be the case, for example, in areas along the border of adjacent nations.

local vertical datums

Even within a country, heights may differ depending on the location of the tide gauge to which mean sea level is related. As an example, the mean sea level from the Atlantic to the Pacific coast of the U.S.A. differs by 0.6–0.7 m. The tide gauge (zero height) of the Netherlands differs -2.34 m from the tide gauge (zero height) of neighbouring Belgium.

The local vertical datum is implemented through a levelling network (Figure 3.3a), which consists of benchmarks whose height above mean sea level has been determined through geodetic levelling . The implementation of the datum enables easy user access. Surveyors, for example, do not need to start from scratch (i.e. from the Amsterdam tide gauge) each time they need to determine the height of a new point. They use the benchmark of the levelling network that is closest to the new point (Figure 3.3b).
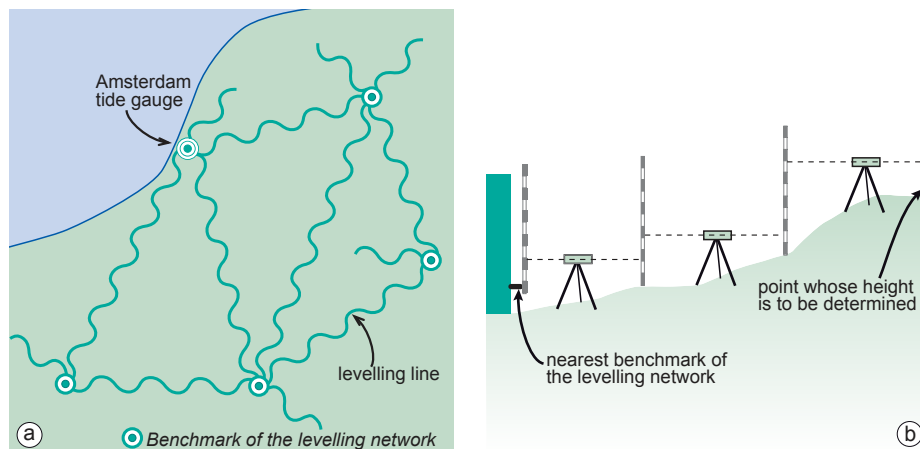
Figure 3.3
A levelling network implements a local vertical datum: (a) network of levelling lines starting from the Amsterdam Tide Gauge, showing some of the benchmarks; (b) how the orthometric height (H) is determined for some point by working from the nearest benchmark.

As a result of satellite gravity missions, it is currently possible to determine height (H) above the Geoid to centimetre levels of accuracy. It is foreseeable that a global vertical datum may become ubiquitous in the next 10–15 years. If all geodata, for example maps, were to use such a global vertical datum, heights would become globally comparable, effectively making local vertical datums redundant for users of geoinformation.

The ellipsoid

We have defined a physical surface, the Geoid, as a reference surface for heights. We also need, however, a reference surface for the description of the *horizontal coordinates* of points of interest. Since we will later want to project these horizontal coordinates onto a mapping plane, the reference surface for horizontal coordinates requires a mathematical definition and description. The most convenient geometric reference

horizontal coordinates

is the *oblate ellipsoid* (Figure 3.4). It provides a relatively simple figure that fits the Geoid to a first-order approximation (for small-scale mapping purposes we may use the *sphere*). An ellipsoid is formed when an ellipse is rotated around its minor axis. This ellipse, which defines an ellipsoid or *spheroid*, is called a meridian ellipse (notice that ellipsoid and spheroid are used here to refer to the same).
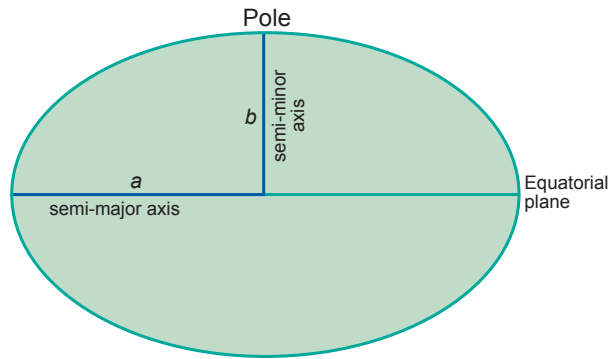
Figure 3.4
An oblate ellipsoid, defined by its semi-major axis a and semi-minor axis b.

The shape of an ellipsoid may be defined in a number of ways, but in geodetic practice it is is usually defined by its semi-major axis and flattening (Figure 3.4). Flattening f is dependent on both the semi-major axis a and the semi-minor axis b:

$$f = \frac{a - b}{a}.$$

The ellipsoid may also be defined by its semi-major axis a and its eccentricity e, which can be expressed as:

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

Given one axis and any one of the other three parameters, the other two can be derived. Typical values of the parameters for an ellipsoid are:

$$a = 6378135.00 \; m, \; b = 6356750.52 \; m, \; f = \frac{1}{298.26}, \; e = 0.08181881066$$

local ellipsoids

Many different sorts of ellipsoids have been defined. Local ellipsoids have been established to fit the Geoid (mean sea level) well over an area of local interest, which in the past was never larger than a continent. This meant that the differences between the Geoid and the reference ellipsoid could effectively be ignored, allowing accurate maps to be drawn in the vicinity of the datum (Figure 3.5).

global ellipsoids

With increasing demands for global surveying, work is underway to develop global reference ellipsoids. In contrast to local ellipsoids, which apply only to a specific country or localized area of the Earth's surface, global ellipsoids approximate the Geoid as a mean Earth ellipsoid. The International Union for Geodesy and Geophysics (IUGG) plays a central role in establishing these reference figures.

In 1924, the general assembly of the IUGG in Madrid introduced the ellipsoid determined by Hayford in 1909 as the international ellipsoid. According to subsequently acquired knowledge, however, the values for this ellipsoid give an insufficiently accurate approximation. At the 1967 general assembly of the IUGG in Luzern, the 1924 reference system was replaced by the Geodetic Reference System 1967 (GRS 1967) el-

lipsoid. Later, in 1980, this was in turn replaced by the Geodetic Reference System 1980 (GRS80) ellipsoid.

Name	*a* (m)	*b* (m)	*f*
International (1924)	6378388.	6356912.	1:297.000
GRS 1967	6378160.	6356775.	1:298.247
GRS 1980 & WGS84	6378137.	6356752.	1:298.257

The local horizontal datum

Ellipsoids have varying positions and orientations. An ellipsoid is positioned and oriented with respect to the local mean sea level by adopting a latitude (ϕ) and longitude (λ) and ellipsoidal height (h) of what is called a fundamental point and an azimuth to an additional point. We say that this defines a *local horizontal datum*. Note that the term horizontal datum and geodetic datum are treated as equivalent and interchangeable terms.

Several hundred local horizontal datums exist in the world. The reason for this is obvious: different local ellipsoids of varying position and orientation had to be adopted to provide a best fit of the local mean sea level in different countries or regions. The Potsdam Datum, the local horizontal datum used in Germany is an example of a local horizontal datum. The fundamental point is located in Rauenberg and the underlying ellipsoid is the Bessel ellipsoid ($a = 6,377,397.156$ m, $b = 6,356,079.175$ m). We can determine the latitude and longitude (ϕ, λ) of any other point in Germany with respect to this local horizontal datum using geodetic positioning techniques, such as triangulation and trilateration. The result of this process will be the geographic (or horizontal) coordinates (ϕ, λ) of a new point in the Potsdam Datum.

A local horizontal datum is determined through a triangulation network. Such a network consists of monumented points that form a network of triangular mesh elements (Figure 3.6). The angles in each triangle are measured, in addition to at least one side of the triangle; the fundamental point is also a point in the triangulation network. The angle measurements and the adopted coordinates of the fundamental point are then used to derive geographic coordinates (ϕ, λ) for all monumented points of the triangulation network.

triangulation networks

Within this framework, users do not need to start from scratch (i.e. from the fundamental point) in order to determine the geographic coordinates of a new point. They

can use the monument of the triangulation network that is closest to the new point. The extension and re-measurement of the network is nowadays done through satellite measurements.
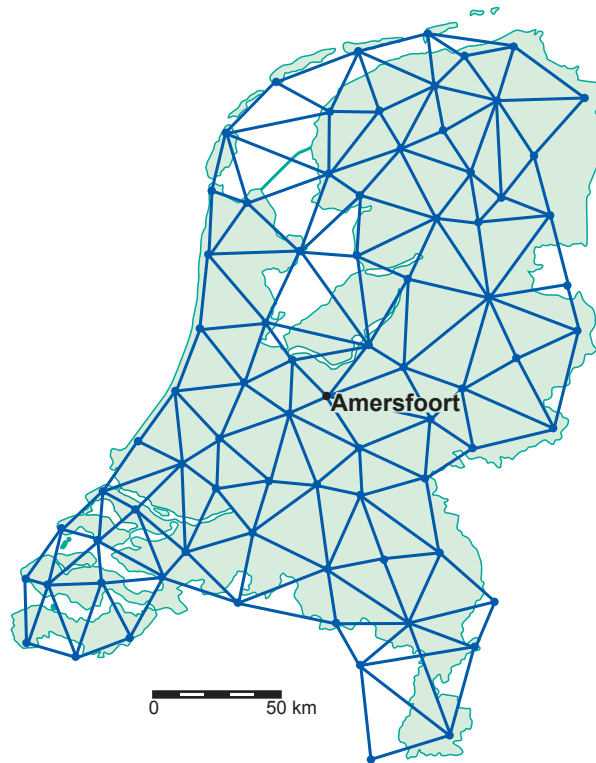
Figure 3.6
The old primary triangulation network in the Netherlands was made up of 77 points (mostly church towers). The extension and re-measurement of the network is done nowadays through satellite measurements. Adapted from original figure by "Dutch Cadastre and Land Registers", now called *het Kadaster*.

The global horizontal datum

With increasing demands for global surveying, activities are underway to establish global reference surfaces. The motivation in this is to make geodetic results mutually compatible and to be able to provide coherent results to other disciplines, e.g. astronomy and geophysics.

ITRS

The most important global (geocentric) spatial reference system for the geoinformation community is the International Terrestrial Reference System (ITRS) . This is a three-dimensional coordinate system with a well-defined origin (the centre of mass of the Earth) and three orthogonal coordinate axes (X, Y, Z). The Z-axis points towards a mean North Pole. The X-axis is oriented towards the mean Greenwich meridian and is orthogonal to the Z-axis. The Y-axis completes the right-handed reference coordinate system (Figure 3.7a).

ITRF

The ITRS is realized through the International Terrestrial Reference Frame (ITRF), a distributed set of ground control stations that measure their position continuously using GPS (Figure 3.7b). Constant re-measuring is needed because of the addition of new control stations and ongoing geophysical processes (mainly tectonic plate motion) that deform the Earth's crust at measurable global, regional and local scales. These deformations cause positional differences over time and have resulted in more than one realization of the ITRS. Examples are the ITRF96 and the ITRF2000. The ITRF96 was established on 1 January 1997, which means that the measurements use data acquired up to 1996 to fix the geocentric coordinates (X, Y and Z in metres) and velocities (posi-

tional change in X, Y and Z in metres per year) of the different stations. The velocities are used to propagate measurements to other epochs (times). The trend is to use the ITRF everywhere in the world for reasons of global compatibility.

GPS uses the World Geodetic System 1984 (WGS84) as its reference system. It has been refined on several occasions and is now aligned with the ITRF to within a few centimetres worldwide. Global horizontal datums, such as ITRF2000 or WGS84, are also called geocentric datums because they are geocentrically positioned with respect to the centre of mass of the Earth. They became available only recently (roughly, since the 1960s), as a result of advances in extra-terrestrial positioning techniques. [1]

geocentric datums

Since the range and shape of satellite orbits are directly related to the centre of mass of the Earth, observations of natural or artificial satellites can be used to pinpoint the centre of mass of the Earth, and hence the origin of the ITRS[2]. This technique can also be used for the realization of global ellipsoids and datums at levels of accuracy required for large-scale mapping.

To implement the ITRF in a particular region, a densification of control stations is needed to ensure that there are enough coordinated reference points available in that region. These control stations are equipped with permanently operating satellite positioning equipment (i.e. GPS receivers and auxiliary equipment) and communication links. Examples of (networks consisting of) such permanent tracking stations are the Actief GNSS Referentie Systeem Nederland (AGRS) in the Netherlands and the Satellitenpositionierungsdienst der deutschen Landesvermessung (SAPOS) in Germany.

We can transform ITRF coordinates (X, Y and Z in metres) into geographic coordinates (ϕ, λ, h) with respect to the GRS80 ellipsoid without the loss of accuracy. However the ellipsoidal height h obtained through this straightforward transformation has no physical meaning and is contrary to our intuitive human perception of height. Therefore, we use instead the height, H, above the Geoid (see Figure 3.8). It is foreseeable that global 3D spatial referencing in terms of (ϕ, λ, H) could become ubiquitous in the next 10–15 years. If, by then, all published maps are also globally referenced the underlying spatial referencing concepts will become transparent and, hence, irrelevant to users of geoinformation.

3D spatial referencing

---

[1]Extra-terrestrial positioning techniques include, for example, Satellite Laser Ranging (SLR), Lunar Laser Ranging (LLR), Global Positioning System (GPS), and Very Long Baseline Interferometry (VLBI).

[2]In the case of an idealized spherical Earth, it is one of the focal points of the elliptical orbits.

Figure 3.8
Height h above the geocentric ellipsoid, and height H above the Geoid. h is measured orthogonal to the ellipsoid, H orthogonal to the Geoid.
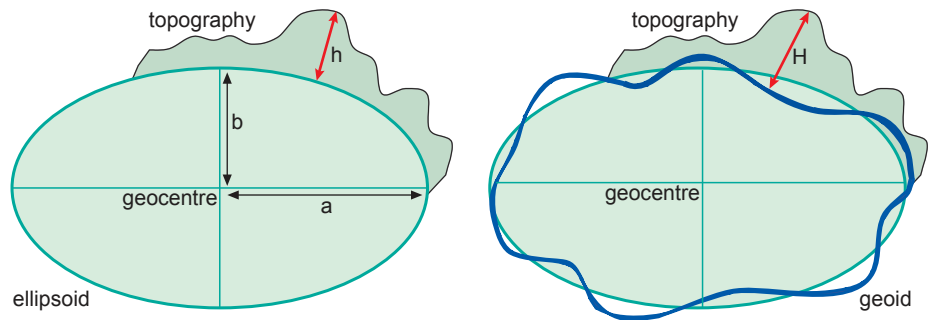
Hundreds of existing local horizontal and vertical datums are still relevant because they form the basis of map products all over the world. For the next few years we still have to deal with both local and global datums, until the former are eventually phased out. During the transition period, we will need tools to transform coordinates from local horizontal datums to a global horizontal datum and vice versa (see Subsection 3.1.4). The organizations that usually develop transformation tools and make them available to the user community are provincial or national mapping organizations (NMOs) and cadastral authorities.

3.1.2 Coordinate systems

Spatial data are special, because they are spatially referenced. Different kinds of coordinate systems are used to position data in space. Here we distinguish between *spatial* and *planar* coordinate systems. *Spatial* (or global) coordinate systems locate data either on the Earth's surface in a 3D space or on the Earth's reference surface (ellipsoid or sphere) in a 2D space. *Planar* coordinate systems, on the other hand, locate data on the flat surface of a map in a 2D space. Initially the 2D Cartesian coordinate system and the 2D polar coordinate system will be examined. This will be followed by a discussion of the geographic coordinate system in a 2D and 3D space and the geocentric coordinate system, also known as the 3D Cartesian coordinate system.

spatial coordinate systems

planar coordinate systems

2D geographic coordinates (ϕ, λ)

The most widely used global coordinate system consists of lines of geographic *latitude* (phi or ϕ or φ) and *longitude* (lambda or λ). Lines of equal latitude are called parallels. They form circles on the surface of the ellipsoid.[3]. Lines of equal longitude are called meridians and form ellipses (meridian ellipses) on the ellipsoid (Figure 3.9)

The latitude (ϕ) of a point P (Figure 3.10) is the angle between the ellipsoidal normal through P' and the equatorial plane. Latitude is zero on the Equator ($\phi = 0°$), and increases towards the two poles to maximum values of $\phi = +90°$ (N $90°$) at the North Pole and $\phi = -90°$ (S $90°$) at the South Pole.

The longitude (λ) of the point is the angle between the meridian ellipse that passes through Greenwich and the meridian ellipse containing the point in question. It is measured on the equatorial plane from the meridian of Greenwich ($\lambda = 0°$), either eastwards through $\lambda = +180°$ (E $180°$) or westwards through $\lambda = -180°$ (W $180°$).

Latitude and longitude represent the geographic coordinates (ϕ, λ) of a point P' (Figure 3.10) with respect to the selected reference surface. They are always given in angular units. For example, the coordinates for the City Hall in Enschede are:[4]

---

[3]The concept of geographic coordinates can also be applied to a sphere.
[4]This latitude and longitude refers to the Amersfoort datum. The use of a different reference surface will
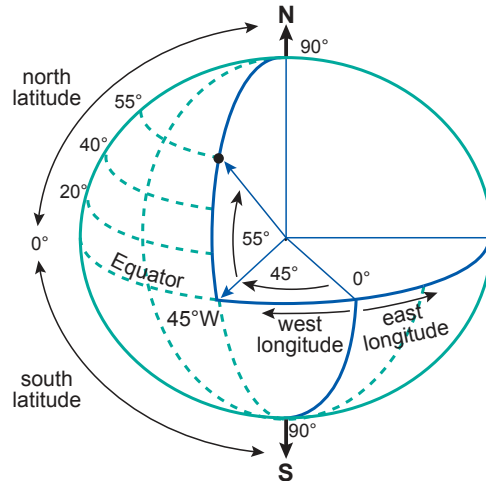
$$\phi = 52°13'26.2''N, \ \lambda = 6°53'32.1''E$$

The graticule on a map represents the projected position of the geographic coordinates (ϕ, λ) at constant intervals or, in other words, the projected position of selected meridians and parallels (Figure 3.13). The shape of the graticule depends largely on the characteristics of the map projection and the scale of the map.

3D geographic coordinates (ϕ, λ, h)

3D geographic coordinates (ϕ, λ, h) are obtained by introducing ellipsoidal height (h) into the system. The ellipsoidal height (h) of a point is the vertical distance of the point in question above the ellipsoid. It is measured in distance units along the ellipsoidal normal from the point to the ellipsoid surface. 3D geographic coordinates can be used to define a position on the surface of the Earth (point P in Figure 3.10).
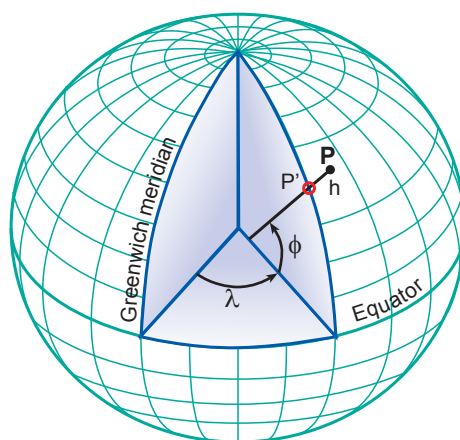
Figure 3.10
The angles of latitude (ϕ) and
longitude (λ) and the
ellipsoidal height (h)
represent the 3D geographic
coordinate system.

result in different angles of latitude and longitude.

3D geocentric coordinates (X, Y, Z)

An alternative method for defining a 3D position on the surface of the Earth is to use geocentric coordinates (X, Y, Z), also known as *3D Cartesian coordinates*. The system's origin lies at the Earth's centre of mass, with the X and Y axes on the plane of the Equator. The X-axis passes through the meridian of Greenwich and the Z-axis coincides with the Earth's axis of rotation. The three axes are mutually orthogonal and form a right-handed system. Geocentric coordinates can be used to define a position on the surface of the Earth (point P in Figure 3.11).

polar motion

The rotational axis of the Earth, however, changes position over time (referred to as *polar motion*). To compensate for this, the mean position of the pole in the year 1903 (based on observations between 1900 and 1905) is used to define what is referred to as the "Conventional International Origin" (CIO).
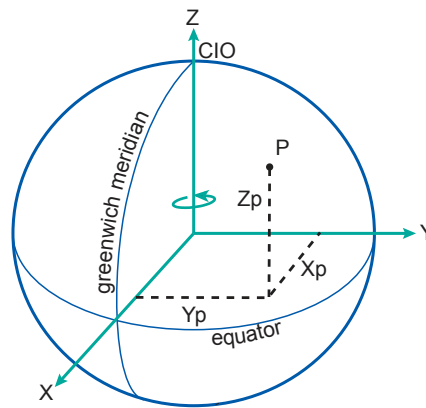
Figure 3.11
An illustration of the 3D geocentric coordinate system (see text for further explanation).

2D Cartesian coordinates (X, Y)

A flat map has only two dimensions: width (left to right) and length (bottom to top). Transforming the three dimensional Earth onto a two-dimensional map is the subject matter of map projections and coordinate transformations (Subsection 3.1.3 and Subsection 3.1.4). Here, as for several other cartographic applications, *two-dimensional Cartesian coordinates* (x, y), also known as *planar rectangular coordinates*, describe the location of any point unambiguously.

The 2D Cartesian coordinate system is one of intersecting perpendicular lines with the X-axis and the Y-axis as principal axes. The X-axis (the *Easting*) is the horizontal axis and the Y-axis (the *Northing*) is the vertical axis with an intersection at the *origin*. The plane is marked at intervals by equally-spaced coordinate lines that together form the *map grid* . Given two numerical coordinates x and y for point P, one can unambiguously specify any location P on the map (Figure 3.12).

false origin

Usually, the origin is assigned the coordinates $x = 0$ and $y = 0$. Sometimes, however, large positive values are added to the origin coordinates. This is to avoid negative values for the x and y coordinates in cases where the origin of the coordinate system is located inside the specific area of interest. The point that then has the coordinates $x = 0$ and $y = 0$ is called the *false origin*. The Rijksdriehoekstelsel (RD) in the Netherlands is an example of a system with a false origin. The system is based on the azimuthal double stereographic projection (see Section 3.1.3), with the Bessel ellipsoid used as reference surface. The origin was shifted from the projection centre (Amersfoort) towards the southwest(false origin)to avoid negative coordinates inside the country (see Figure 3.13).
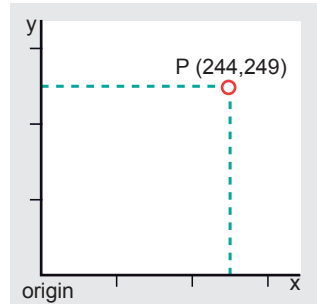
The grid on a map represents lines having constant 2D Cartesian coordinates (Figure 3.13). It is almost always a rectangular system and is used on large- and medium-scale maps to enable detailed calculations and positioning. The map grid is usually not used on small-scale maps (about 1:1,000,000 or smaller). Scale distortions that result from transforming the Earth's curved surface to the mapping plane are so great on small-scale maps that detailed calculations and positioning become difficult.

map grid

2D Polar coordinates (α, d)

Another way of defining a point in a plane is by using polar coordinates. This is the distance d from the origin to the point concerned and the angle α between a fixed (or zero) direction and the direction to the point. The angle α is called *azimuth* or *bearing*

103

and is measured in a clockwise direction. It is given in angular units while the distance d is expressed in length units.
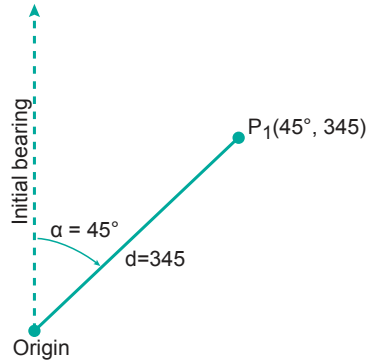
Figure 3.14
An illustration of the 2D Polar coordinate system (see text for further explanation).

Bearings are always related to a fixed direction (initial bearing) or a datum line. In principle, this reference line can be chosen freely. Three different, widely used fixed directions are: *True North*, *Grid North* and *Magnetic North*. The corresponding bearings are true (or geodetic) bearings, grid bearings and magnetic (or compass) bearings, respectively.

polar coordinates

Polar coordinates are often used in land surveying. For some types of surveying instruments, it is advantageous to make use of this coordinate system. The development of precise, remote-distance measurement techniques has led to a virtually universal preference for the polar coordinate method for detailed surveys.

3.1.3 Map projections

Maps are one of the world's oldest types of document. In the days that our planet was thought to be *flat*, a map was simply a miniature representation of a part of the world. To represent the specifically curved Earth's surface, a map needs to be a flattened representation of a part of the planet. Map projection concerns itself with ways of translating the curved surface of the Earth into a flat, 2D map.

> *Map projection* is a mathematically described technique for representing the Earth's curved surface on a flat map.

To represent parts of the surface of the Earth on a flat, printed map or a computer screen, the curved horizontal reference surface must be mapped onto a 2D mapping plane. The reference surface for large-scale mapping is usually an oblate ellipsoid; for small-scale mapping it is a sphere.[5] Mapping onto a 2D mapping plane means transforming each point on the reference surface with geographic coordinates (ϕ, λ) to a set of Cartesian coordinates (x, y) that represent positions on the map plane (Figure 3.15).

mapping equations

The actual mapping cannot usually be visualized as a true geometric projection, directly onto the mapping plane. Rather, it is achieved through mapping equations. A *forward mapping equation* transforms the geographic coordinates (ϕ, λ) of a point on the curved reference surface to a set of planar Cartesian coordinates (x, y), representing the position of the same point on the map plane:

$$(x, y) = f(\phi, \lambda)$$

---

[5]In practice, maps at scales of 1:1,000,000 or smaller can use the mathematically simpler sphere without the risk of large distortions. At larger scales, the more complicated mathematics of ellipsoids is needed to prevent large distortions occurring on the map.

Map plane

Y

P(x,y)

P'(ϕ,λ)

X

ϕ

λ

Reference surface

The corresponding *inverse mapping equation* transforms mathematically the planar Cartesian coordinates (x, y) of a point on the map plane to a set of geographic coordinates (ϕ, λ) on the curved reference surface:

$$(\phi, \lambda) = f(x,\ y)$$

The Mercator projection (spherical assumption) [106], a commonly used mapping projection, can be used to illustrate the use of mapping equations. The *forward mapping equation* for the Mercator projection is:[6]

$$x = R(\lambda - \lambda_0)$$
$$y = R \ln \tan \left(\frac{\pi}{4} + \frac{\phi}{2} \right)$$

The *inverse mapping equation* for the Mercator projection is:

$$\phi = \frac{\pi}{2} - 2 \arctan \left(e^{-\frac{y}{R}} \right)$$
$$\lambda = \frac{x}{R} + \lambda_0$$

Classification of map projections

Many map projections have been developed, each with its own specific qualities. It is these qualities that make the resulting maps useful for certain purposes. By definition, any map projection is associated with scale distortions. There is simply no way to flatten an ellipsoidal or spherical surface without stretching some parts of the surface more than others. The amount and kind of distortions a map has depends on the type of map projection.

scale distortions

Some map projections can be visualized as true geometric projections directly onto the mapping plane—known as an azimuthal projections—or onto an intermediate surface,

---

[6]When an ellipsoid is used as a reference surface, the equations are considerably more complicated than those introduced here. R is the radius of the spherical reference surface at the scale of the map; ϕ and λ are given in radians; λ_0 is the central meridian of the projection; $e = 2.7182818$, the base of the natural logarithms, not the eccentricity.

which is then rolled out onto the mapping plane. Typical choices for such intermediate surfaces are cones and cylinders. These map projections are called conical or cylindrical projections, respectively. Figure 3.16 shows the surfaces involved in these three classes of projection.
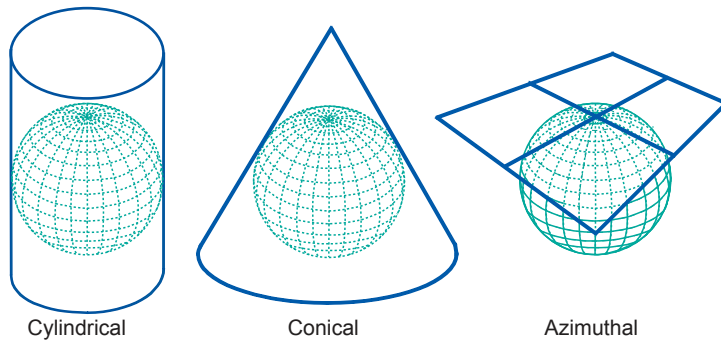
Figure 3.16
Classes of map projections

Cylindrical   Conical   Azimuthal

The azimuthal, conical, and cylindrical surfaces in Figure 3.16 are all *tangent* surfaces, i.e. they touch the horizontal reference surface at one point (azimuthal), or along a closed line (cone and cylinder), only. Another class of projections is obtained if the surfaces are chosen to be *secant* to (to intersect with) the horizontal reference surface; see Figure 3.17. Then, the reference surface is intersected along one closed line (azimuthal) or two closed lines (cone and cylinder). Secant map surfaces are used to reduce or average out scale errors since the line(s) of intersection are not distorted on the map.
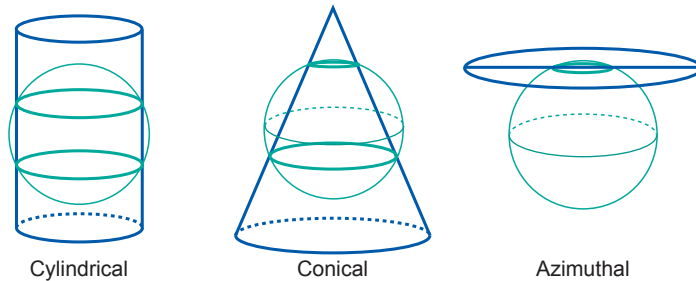
Figure 3.17
Three classes of secant projection

Cylindrical   Conical   Azimuthal

In the geometric depiction of map projections in Figures 3.16 and 3.17, the symmetry axes of the plane, cone and cylinder coincide with the rotation axis of the ellipsoid or sphere, i.e. a line through the North and South poles. In this case, the projection is said to be a *normal projection*. The other cases are *transverse projections* (symmetry axis in the equatorial plane) and *oblique projections* (symmetry axis is somewhere between the rotation axis and the equator of the ellipsoid or sphere); see Figure 3.18.

normal, transverse, and oblique projections

The Universal Transverse Mercator (UTM) is a system of map projection that is used worldwide. It is derived from the Transverse Mercator projection (also known as Gauss-Kruger or Gauss conformal projection). UTM uses a transverse cylinder secant to the horizontal reference surface. It divides the world into 60 narrow longitudinal zones of 6 degrees, numbered from 1 to 60. The narrow zones of 6 degrees (and the secant map surface) make the distortions small enough for large-scale topographic mapping.

Normal cylindrical projections are typically used to map the world in its entirety. Conical projections are often used to map individual continents, whereas the normal az-
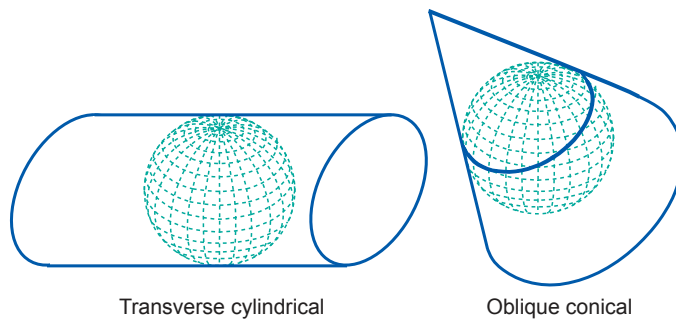
Transverse cylindrical   Oblique conical

imuthal projection may be used to map polar areas. Transverse and oblique aspects of many projections can be used for most parts of the world.

It is also important to consider the shape of the area to be mapped. Ideally, the general shape of the mapping area should be well-match with the distortion pattern of a specific projection. If an area is approximately circular, it is possible to create a map that minimizes distortion for that area on the basis of an azimuthal projection. Cylindrical projection is best for a rectangular area and conic projection for a triangular area.

So far, we have not specified *how* the curved horizontal reference surface is projected onto a plane, cone or cylinder. *How* this is done determines what kind of *distortions* the map will have compared to the original curved reference surface. The distortion properties of a map are typically classified according to what is *not* distorted on the map:

distortion properties

- With a *conformal* map projection, the angles between lines in the map are identical to the angles between the original lines on the curved reference surface. This means that angles (with short sides) and shapes (of small areas) are shown correctly on the map.

- With an *equal-area* (equivalent) map projection, the areas in the map are identical to the areas on the curved reference surface (taking into account the map scale), which means that areas are represented correctly on the map.

- With an *equidistant* map projection, the length of particular lines in the map are the same as the length of the original lines on the curved reference surface (taking into account the map scale).

A particular map projection can exhibit only one of these three properties. No map projection can be both conformal and equal-area, for example.

The most appropriate type of distortion for a map depends largely on the purposes for which the map will be used. Conformal map projections represent angles correctly, but as the region becomes larger they show considerable area distortions (Figure 3.19). Maps used for the measurement of angles (e.g. aeronautical charts, topographic maps) often make use of a conformal map projection such as the UTM projection.

Equal-area projections, on the other hand, represent areas correctly, but as the region becomes larger, considerable distortions of angles and, consequently, shapes occur (Figure 3.20). Maps that are to be used for measuring area (e.g. distribution maps) are often made using an equal-area map projection.

The equidistant property is achievable only to a limited degree. That is, true distances can be shown only from one or two points to any other point on the map, or in certain directions. If a map is true to scale along the meridians (i.e. no distortion in the

Figure 3.19
The Mercator projection, a cylindrical map projection with conformal properties. The area distortions are significant towards the polar regions.

Figure 3.20
The cylindrical equal-area projection, i.e. a cylindrical map projection with equal-area properties. Distortions of shapes are significant towards the poles.

North–South direction), we say that the map is *equidistant along the meridians* (e.g. an equidistant cylindrical projection) (Figure 3.21). If a map is true to scale along all parallels we say the map is *equidistant along the parallels* (i.e. no distortion in the East–West direction). Maps for which the area and angle distortions need to be reasonably acceptable (several thematic maps) often make use of an equidistant map projection.

As these discussions indicate, a particular map projection can be classified. An example would be the classification "conformal conic projection with two standard parallels", which means that the projection is a conformal map projection, that the intermediate surface is a cone, and that the cone intersects the ellipsoid (or sphere) along two parallels. In other words, the cone is secant and the cone's symmetry axis is parallel to the rotation axis. (This would amount to the middle projection displayed in Figure 3.17). This projection is also referred to as "Lambert's conical projection" [47].

3.1.4   Coordinate transformations

Users of geoinformation often need transformations from a particular 2D coordinate system to another system. This includes the transformation of polar coordinates into Cartesian map coordinates, or the transformation from one 2D Cartesian (x, y) system of a specific map projection into another 2D Cartesian (x', y') system of a defined map projection. This transformation is based on relating the two systems on the basis of a set of selected points whose coordinates are known in both systems, such as ground control points or common points such as corners of houses or road intersections. Image and scanned data are usually transformed by this method. The transformations may be conformal, affine, polynomial or of another type, depending on the geometric errors in the data set.

2D Polar to 2D Cartesian transformations

The transformation of polar coordinates (α, d), into Cartesian map coordinates (x, y) is done when field measurements, i.e. angular and distance measurements, are transformed into map coordinates. The equation for this transformation is:

$$x = d \sin \alpha$$

$$y = d \cos \alpha$$

The inverse equation is:

$$\alpha = \arctan \left(\frac{x}{y} \right)$$

$$d^2 = x^2 + y^2$$

Changing map projection

Forward and inverse mapping equations are normally used to transform data from one map projection into another. The inverse equation of the source projection is used first to transform source projection coordinates (x, y) to geographic coordinates (ϕ, λ). Next, the forward equation of the target projection is used to transform the geographic coordinates (ϕ, λ) into target projection coordinates (x', y'). The first equation takes us from a projection A into geographic coordinates. The second takes us from geographic

coordinates (ϕ, λ) to another map projection B. These principles are illustrated in Figure 3.22.

Historically, GI Science has dealt with data referenced spatially with respect to the (x, y) coordinates of a specific map projection. For application domains requiring 3D spatial referencing, a height coordinate may be added to the (x, y) coordinates of the point. The additional height coordinate can be a height H above mean sea level, which is a height with a physical meaning. The (x, y, H) coordinates then represent the location of objects in a 3D space.
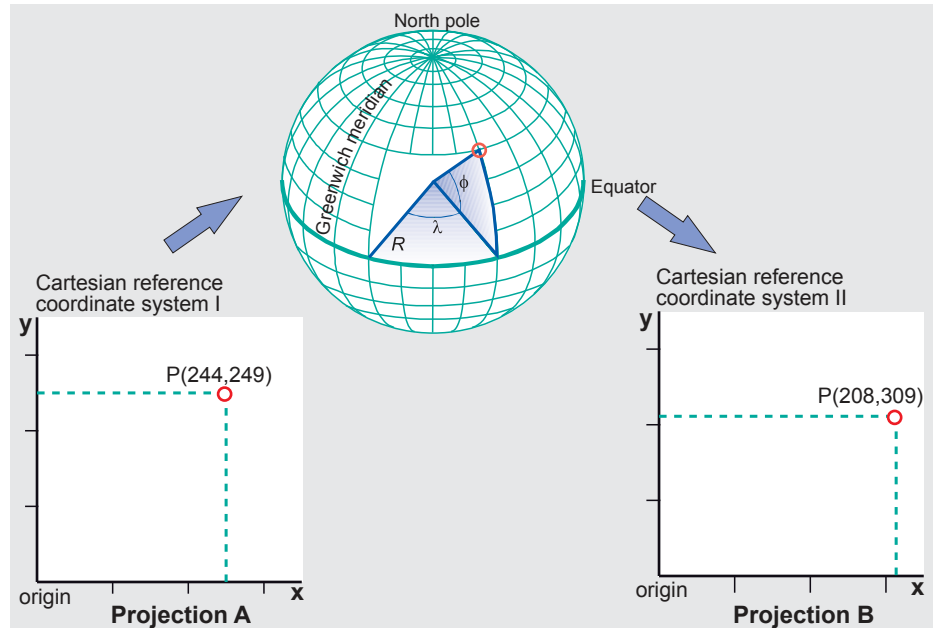
Figure 3.22
The principle of changing from one map projection to another.

Datum transformations

A change of map projection may also include a change of the horizontal datum. This is the case when the source projection is based upon a different horizontal datum than the target projection. If the difference in horizontal datums is ignored, there will not be a perfect match between adjacent maps of neighbouring countries or between overlaid maps originating from different projections. It may lead to differences of several hundreds of metres in the resulting coordinates. Therefore, spatial data with different underlying horizontal datums may require *datum transformation*.

Suppose we wish to transform spatial data from the UTM projection to the Dutch RD system, and suppose that the data in the UTM system are related to the European Datum 1950 (ED50), while the Dutch RD system is based on the Amersfoort datum. To achieve a perfect match, in this example the change of map projection should be combined with a datum transformation step; see Figure 3.23.

The inverse equation of projection A is used first to take us from the map coordinates (x, y) of projection A to the geographic coordinates (ϕ, λ, h) for datum A. A height coordinate (h or H) may be added to the (x, y) map coordinates. Next, the datum transformation takes us from these coordinates to the geographic coordinates (ϕ, λ, h) for datum B. Finally, the forward equation of projection B takes us from the geographic coordinates (ϕ, λ, h) for datum B to the map coordinates (x', y') of projection B.

Mathematically, a datum transformation is feasible via the geocentric coordinates (x, y, z)

or directly by relating the geographic coordinates of both datum systems. The latter relates the ellipsoidal latitude (ϕ) and longitude (λ), and possibly also the ellipsoidal height (h), of both datum systems [59].

Geographic coordinates (ϕ, λ, h) can be transformed into geocentric coordinates (x, y, z), and vice versa. The datum transformation via the geocentric coordinates implies a 3D similarity transformation. This is essentially a transformation between two orthogonal 3D Cartesian spatial reference frames together with some elementary tools from adjustment theory. The transformation is usually expressed with seven parameters: three rotation angles (α, β, γ), three origin shifts (X_0, Y_0, Z_0) and a scale factor (s). The inputs are the coordinates of points in datum A and coordinates of the same points in datum B. The output are estimates of the seven transformation parameters and a measure of the likely error of the estimate.

Datum transformation parameters have to be estimated on the basis of a set of selected points whose coordinates are known in both datum systems. If the coordinates of these points are not correct—often the case for points measured on a local datum system—the estimated parameters may be inaccurate and hence the datum transformation will be inaccurate.

Inaccuracies often occur when we transform coordinates from a local horizontal datum to a global geocentric datum. The coordinates in the local horizontal datum may be distorted by several tens of metres because of the inherent inaccuracies of the measurements used in the triangulation network. These inherent inaccuracies are also responsible for another complication: the transformation parameters are not unique. Their estimation depends on the particular choice of common points and whether all

datum transformation
parameters

seven transformation parameters, or only some of them, are estimated.

The example in Table 3.2 illustrates the transformation of the Cartesian coordinates of a point in the state of Baden-Württemberg, Germany, from ITRF to Cartesian coordinates in the Potsdam Datum. Sets of numerical values for the transformation parameters are available from three organizations:

Table 3.2
Three different sets of datum transformation parameters from three different organizations for transforming a point from ITRF to the Potsdam datum.

	Parameter	National set	Provincial set	NIMA set
scale	s	$1 - 8.3 \cdot 10^{-6}$	$1 - 9.2 \cdot 10^{-6}$	1
angles	α	$+1.04''$	$+0.32''$	
	β	$+0.35''$	$+3.18''$	
	γ	$-3.08''$	$-0.91''$	
shifts (m)	X_0	-581.99	-518.19	-635
	Y_0	-105.01	-43.58	-27
	Z_0	-414.00	-466.14	-450

1. The federal mapping organization of Germany (labelled "National set" in Table 3.2) provided a set calculated using common points distributed throughout Germany. This set contains all seven parameters and is valid for whole Germany.

2. The mapping organization of Baden-Württemberg (labelled "Provincial set" in Table 3.2) provided a set calculated using common points distributed throughout the state of Baden-Württemberg. This set contains all seven parameters and is valid only within the state borders.

3. The National Imagery and Mapping Agency (NIMA) of the U.S.A. (labelled "NIMA set" in Table 3.2) provided a set calculated using common points distributed throughout Germany and based on the ITRF. This set contains a coordinate shift only (no rotations, and scale equals unity). This set is valid for whole Germany.

The three sets of transformation parameters vary by several tens of metres, for reasons already mentioned. The sets of transformation parameters were used to transform the ITRF cartesian coordinates of a point in the state of Baden-Württemberg. Its ITRF (X, Y, Z) coordinates are:

$$(4, 156, 939.96 \text{ m}, \ 671, 428.74 \text{ m}, \ 4, 774, 958.21 \text{ m}).$$

The three sets of transformed coordinates for the Potsdam datum are given in Table 3.3.

Table 3.3
Three sets of transformed coordinates for a point in the state of Baden-Württemberg, Germany.

Potsdam coordinates	National set (m)	Provincial set (m)	NIMA set (m)
X	$4, 156, 305.32$	$4, 156, 306.94$	$4, 156, 304.96$
Y	$671, 404.31$	$671, 404.64$	$671, 401.74$
Z	$4, 774, 508.25$	$4, 774, 511.10$	$4, 774, 508.21$

The three sets of transformed coordinates differ by only a few metres from each other. In a different country, the level of agreement could be a within centimetres, but it can be up to tens of metres of each other, depending upon the quality of implementation of the local horizontal datum.

3.2 Satellite-based positioning

The importance of satellites in spatial referencing has already been mentioned before. Satellites have allowed us to create geocentric reference systems and to increase the level of spatial accuracy substantially. Satellite-based systems are critical tools in geodetic engineering for the maintenance of the ITRF. They also play a key role in mapping and surveying in the field, as well as in a growing number of applications requiring positioning techniques. The setting up a satellite-based positioning system requires the implementation of three hardware segments:

1. the *space segment*, i.e. the satellites that orbit the Earth and the radio signals that they emit;

2. the *control segment*, i.e. the ground stations that monitor and maintain the components of the space segment;

3. the *user segment*, i.e. the users, along with the hardware and software they use for positioning.

In satellite positioning, the central problem is to determine the values (X, Y, Z) of a receiver of satellite signals, i.e. to determine the position of the receiver with the accuracy and precision required. The degree of accuracy and precision needed depends on the application, as does timeliness, i.e. are the position values required in real time or can they be determined later during post-processing. Finally, some applications, such as navigation, require kinematic approaches, which take into account the fact that the receiver is not stationary, but moving.

Some fundamental aspects of satellite-based positioning and a brief review of currently available technologies follows.

3.2.1 Absolute positioning

The working principles of absolute, satellite-based positioning are fairly simple:

1. A satellite, equipped with a clock, sends a radio message at a specific moment that includes

   (a) the *satellite identifier*,

   (b) its *position in orbit*, and

   (c) its *clock reading*.

2. A receiver on or above the planet, also equipped with a clock, receives the message slightly later and reads its own clock.

3. From the time delay observed between the two clock readings, and knowing the speed of radio transmission through the medium between (satellite) sender and receiver, the receiver can compute the distance to the sender, also known as the satellite's *pseudorange*. This *pseudorange* is the apparent distance from satellite to receiver, computed from the time delay with which its radio signal is received.

Such a computation places the position of the receiver on a sphere of radius equal to the computed pseudorange (see Figure 3.24a). If, instantaneously, the receiver were to do the same with a message from another satellite positioned elsewhere, the position of the receiver would be placed on another sphere. The intersection of the two spheres,

which have different centres, describes a circle as being the set of possible positions of the receiver (see Figure 3.24b). If a third satellite message is taken into consideration, the three spheres intersect at, at most, two positions, one of which is the actual position of the receiver. In most, if not all practical situations where two positions result, one of them is a highly unlikely position for a signal receiver, thus narrowing down the true position of the receiver. The overall procedure is known as *trilateration*: the determination of a position based on three distances.

trilateration

Figure 3.24
Pseudorange positioning: (a) With just one satellite, the receiver position is somewhere on a sphere, (b) With two satellites, the position is located where the two spheres intersect, i.e. in a circle. Not shown: with three satellites, its position is where the three spheres intersect.
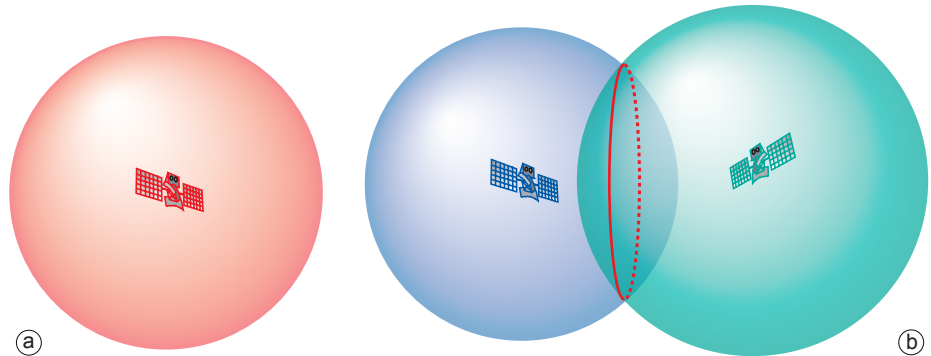
It would appear, therefore, that the signals of three satellites would be sufficient to determine a *positional fix* for our receiver. In theory this is true, but in practice it is not. The reason being that satellite clocks and the receiver clock are never exactly synchronized. Satellite clocks are costly, high-precision, atomic clocks that we can consider synchronized for the time being, but the receiver typically uses a far cheaper, quartz clock that is not synchronized with satellite clocks. This brings an additional unknown variable into play, namely the synchronization bias of the receiver clock, i.e. the difference in time readings between it and the satellite clocks.

clock bias

Our set of unknown variables has now become $(X, Y, Z, \Delta t)$ representing a 3D position and a clock bias. The problem can be solved by including the information obtained from a fourth satellite message, (see Figure 3.25). This will result in the determination of the receiver's actual position (X, Y, Z), as well as its receiver clock bias Δt, and if we correct the receiver clock for this bias we effectively turn it into a high-precision atomic clock as well!

3D positioning

Obtaining a high-precision clock is a fortunate side-effect of using the receiver, as it allows the design of experiments distributed in geographic space that demand high levels of synchronicity. One such application is the use of wireless sensor networks for researching natural phenomena such as earthquakes or meteorological patterns, and for water management.

The positioning of mobile phone users making an emergency call is yet another application. Often callers do not know their location accurately. The telephone company can trace back the call to the receiving transmitter mast, but this may be servicing an area with a radius ranging from 300 m to 6 km. That is far too inaccurate for emergency services. If all masts in the telephony network are equipped with a satellite positioning receiver (and thus, with a very high-precision synchronized clock), however, the time of reception of the call at each mast can be recorded. The *time difference of arrival* of the call between two nearby masts describes a hyperbola on the ground of possible positions of the caller. If the call is received on three masts, two hyperbolas are described, allowing intersection and thus "hyperbolic positioning". With current technology the (horizontal) accuracy would be better than 30 m.

Pseudo Range

Receiver
Clock Bias
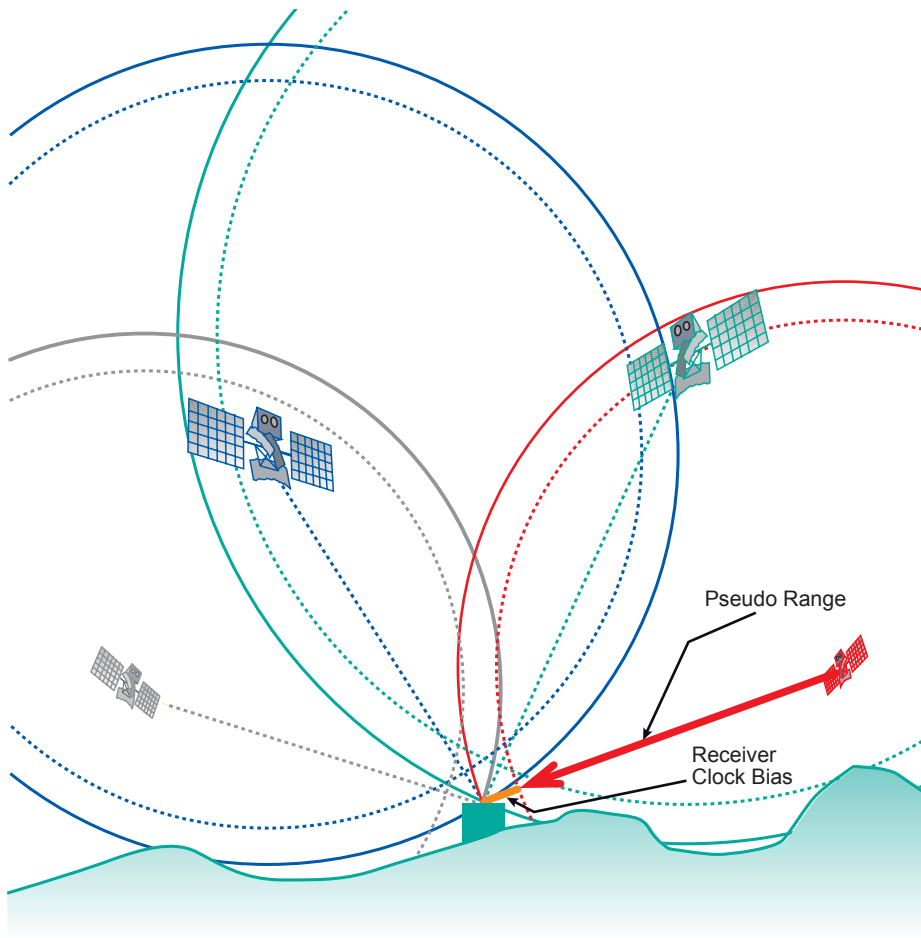
Figure 3.25
Four satellites are needed to
obtain a 3D position fix.
Pseudoranges are indicated
for each satellite as dotted
circles, representing a
sphere; the actual range is
represented as a solid circle,
which is the pseudorange
plus the range error caused
by receiver clock bias.

Returning to satellite-based positioning, when only three, and not four, satellites are "in view", the receiver is capable of falling back from the above *3D positioning mode* to the inferior *2D positioning mode*. With the relative abundance of satellites in orbit around the Earth, this is a relatively rare situation, but it serves to illustrate the importance of 3D positioning.

2D positioning mode

If a 3D fix had already been obtained, the receiver simply assumes that the height above the ellipsoid has not changed since the last 3D fix. If no fix had been obtained, the receiver assumes that it is positioned at the geocentric ellipsoid adopted by the positioning system, i.e. at height $h = 0$.[7] In the receiver computations, the ellipsoid fills the slot of the missing fourth satellite sphere, and the unknown variables can therefore still be determined. Clearly, in both of these cases, the assumption upon which this computation is based is flawed and the resulting positioning in 2D mode will be unreliable—much more so if no previous fix had been obtained and one's receiver is not at all near the surface of the geocentric ellipsoid.

---

[7]Any receiver is capable of transforming a coordinate (X, Y, Z), using a straightforward mathematical transformation, into an equivalent coordinate (ϕ, λ, h), where h is the height above the geocentric ellipsoid.

Time, clocks and world time

Before any notion of standard time existed, villages and cities simply kept track of their local time, determined from the position of the Sun in the sky. When trains became an important means of transportation, these local time systems became problematic as train scheduling required a single time system. Such a time system called for the definition of *time zones*: typically 24 geographic strips bounded by longitudes that are multiples of 15°. This and navigational demands gave rise to Greenwich Mean Time (GMT), based on the mean solar time at the meridian passing through Greenwich, United Kingdom, which is the conventional 0-meridian in geography. GMT became the world time standard of choice.

Greenwich Mean Time

GMT was later replaced by Universal Time (UT), a system still based on meridian crossings of stars, albeit distant quasars, as this approach provides more accuracy than that based on the Sun. It is still the case that the rotational velocity of our planet is not constant and the length of a solar day is increasing. So UT is not a perfect system either. It continues to be used for civilian clock time, but it has now officially been replaced by International Atomic Time (TAI). UT actually has various versions, among them UT0, UT1 and UTC. UT0 is the Earth's rotational time observed at some location. Because the Earth experiences polar motion as well, UT0 differs between locations. If we correct for polar motion, we obtain UT1, which is identical everywhere. Nevertheless, UT1 is still a somewhat erratic clock system because of the varying rotational velocity of the planet, as mentioned above. The degree of uncertainty is about 3 ms per day.

Coordinated Universal Time (UTC) is used in satellite positioning and is maintained with atomic clocks. By convention, it is always within a margin of 0.9 s of UT1, and twice annually it may be shifted to stay within that margin. This occasional shift of a *leap second* is applied at the end of 30 June or, preferably, at the end of 31 December. The last minute of such a day is then either 59 or 61 seconds long. So far, adjustments have always entailed adding a second. UTC time can only be determined to the highest precision after the fact, as atomic time is determined by the reconciliation of the observed differences between a number of atomic clocks maintained by different national time bureaus.

atomic clocks

In recent years, we have learned to measure distance, and therefore also position, with clocks, by using satellite signals, the conversion factor being the speed of light, approximately 3×10^8 m s^{-1} in a vacuum. As a consequence, multiple seconds of clock bias could no longer be accepted, and this is where atomic clocks are at an advantage. They are very accurate time keepers, based on the exact frequencies at which specific atoms (Cesium, Rubidium and Hydrogen) make discrete energy-state jumps. Positioning satellites usually have multiple clocks on board; ground control stations have even better quality atomic clocks.

Atomic clocks are not flawless, however: their timing tends to drift from true time and they, too, need to be corrected. The drift, and the change in drift over time, are monitored and included in the satellite's navigation message, so that these discrepancies can be corrected for.

3.2.2 Errors in absolute positioning

Before we continue discussing other modes of satellite-based positioning, let us take a close look at the potential for error in absolute positioning. Users of receivers are required to be sufficiently familiar with the technology in order to avoid real operating blunders such as poor receiver placement or incorrect receiver software settings, which can render positioning results virtually useless. We will skip over many of the physical and mathematical details underlying these errors; they are only mentioned

here to raise awareness and understanding among users of this technology. For background information on the calculation of positional error (specifically, the calculation of RMSE or *root mean square error*), see Subsection 5.3.2.

Errors related to the space segment

As a first source of error, operators of the control segment may, for example in times of global political tension or war, intentionally deteriorate radio signals from satellites to the general public to avoid optimal use of the system by a perceived enemy. This *selective availability*—meaning that military forces allied with the control segment *will* still have access to undisturbed signals—may cause error that has an order of magnitude larger than all other error sources combined.[8]

A second source occurs if the satellite signal contains incorrect information. Assuming that it will always know its own identifier, the satellite may make two kinds of error:

1. *Incorrect clock reading*. Even atomic clocks can be off by a small margin, and thanks to Einstein we know that moving clocks are slower than stationary clocks, due to a relativistic effect. If one understands that a clock that is off by 0.000001 s causes an computation error in the satellite's pseudorange of approximately 300 m, it becomes clear that these satellite clocks require very strict monitoring.

2. *Incorrect orbit position*. The orbit of a satellite around our planet is easy to describe mathematically if both bodies are considered point masses, but in real life they are not. For the same reasons that the Geoid is not a simply shaped surface, the gravitation pull of the Earth that a satellite experiences in orbit is not simple either. Moreover, satellite orbits are also disturbed by solar and lunar gravitation, making flight paths slightly erratic and difficult to forecast exactly.

Both types of error are strictly monitored by the ground control segment, which is responsible for correcting any errors of this nature, but it does so by applying an agreed-upon tolerance. A control station can obviously compare results of positioning computations such as those discussed above with its accurately *known* position, flagging any unacceptable errors and potentially labelling a satellite as temporarily "unhealthy" until those errors have been corrected and brought back within the agreed tolerance limits. This may be done by uploading a correction to the clock or the satellite's orbit settings.

Errors related to the medium

A third source may be due to the influence of the *medium* between sender and receiver on the satellite's radio signals. The middle atmospheric layers of the stratosphere and mesosphere are relatively harmless and of little hindrance to radio waves, but this is not true of the lower and upper layers of the atmosphere:

- *The troposphere*: the approximate 14 km-high airspace directly above the Earth's surface, which holds most of the atmosphere's oxygen and which envelops all phenomena that we call the weather. It is an obstacle that delays radio waves in a rather variable way.

- *The ionosphere*: the part of the atmosphere that is farthest from the Earth's surface. It starts at an altitude of 90 km and holds many electrically charged atoms,

---

[8]Selective availability was stopped at the beginning of May 2000; in late 2007 the White House decided to remove selective availability capabilities all together. However, when deemed necessary, the US government still has a range of capabilities and technology available to implement regional denial of service of civilian GPS signals in an area of conflict, effectively producing the same result.

thereby forming a protective "shield" against various forms of radiation from space, including, to some extent, radio waves. The degree of ionization shows a distinct night and day rhythm and also varies with solar activity.

The ionosphere is a more severe source of delay for satellite signals, which obviously means that pseudoranges are estimated as being larger than they actually are. When satellites emit radio signals at two or more frequencies, an estimate can be computed from differences in delay incurred for signals of different frequency, which enables correction for atmospheric delay, leading to a 10–50% improvement of accuracy. If this is not the case, or if the receiver is capable of receiving only a single frequency, a model should be applied to forecast the (especially ionospheric) delay; typically the model takes into account the time of day and current latitude of the receiver.

Errors related to the receiver's environment

Fourth in the list of sources of error is that which occurs when a radio signal is received via two or more paths between sender and receiver, typically caused by the signal bouncing off some nearby surface such as a building or rock face. The term applied to this phenomenon is *multi-path*; when it occurs the multiple receptions of the same signal may interfere with each other (see Figure 3.26). Multi-path is a source of error that is difficult to avoid.

multi-path error
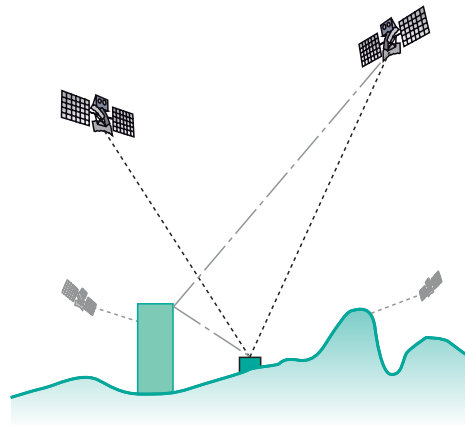
Figure 3.26
At any point in time, a number of satellites will be above the receiver's horizon. But not all of them will be "in view" (e.g. the satellites on the far left and right); and for others, multi-path signal reception may occur.

All of the above sources of error influence computation of a satellite's pseudorange. Cumulatively, they are called the *user equivalent range error* (UERE). Some error sources may affect all satellites being used by a particular receiver, e.g. selective availability and atmospheric delay, while others may be specific to one satellite, for instance, incorrect satellite information and multi-path.

range error

Errors related to the relative geometry of satellites and receiver

There is one more source of error, which is unrelated to individual radio signal characteristics: rather, this error is the result of the combination of signals from satellites used for positioning. The constellation of satellites in the sky from the receiver's perspective is the controlling factor in these cases. Referring to Figure 3.27, the sphere-intersection technique of positioning provides more precise results when the four satellites are evenly spread over the sky; the satellite constellation of Figure 3.27b is preferred over that of 3.27a. This source of error is know as geometric dilution of precision (GDOP). GDOP is lower when satellites are just above the horizon in mutually opposed compass directions. However, such satellite positions have bad atmospheric delay char-

geometric dilution of precision

acteristics, so in practice it is better if they are at least 15° above the horizon. When more than four satellites are in view, modern receivers use "least-squares" adjustment to calculate the best possible positional fix from all the signals. This gives a better solution that obtained just using the "best four", as was done previously.
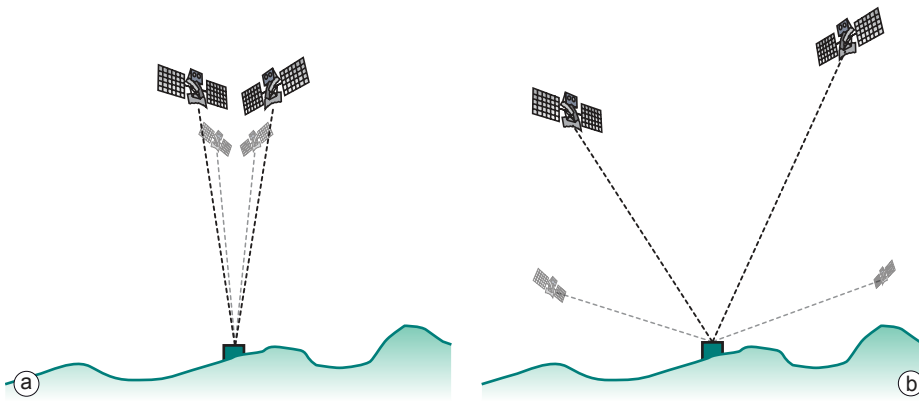
Figure 3.27
Geometric dilution of precision. The four satellites can be in a poor constellation for positioning(a) or in a better constellation (b).

satellite clock (m)	2
satellite position (m)	2.5
ionospheric delay (m)	5
tropospheric delay (m)	0.5
receiver noise (m)	0.3
multi-path (m)	0.5
Total RMSE Range error (m):	
$\sqrt{2^2 + 2.5^2 + 5^2 + 0.5^2 + 0.3^2 + 0.5^2} =$	5.97

Table 3.4
Indication of typical magnitudes of error in absolute satellite-based positioning

These errors are not all of similar magnitude. An overview of some typical values (without selective availability) is provided in Table 3.4. GDOP functions not so much as an independent error source but rather as a multiplying factor, decreasing the precision of position and time values obtained.

The procedure that we discussed above is known as *absolute, single-point positioning based on code measurement*. It is the fastest and simplest, yet least accurate, means of determining a position using satellites. It suffices for recreational purposes and other applications that require horizontal accuracies to within 5–10 m. Typically, when encrypted military signals can also be used, on a dual-frequency receiver the achievable horizontal accuracy is 2–5 m. Below, we discuss other satellite-based positioning techniques with better accuracies.

3.2.3 Relative positioning

One technique for trying to remove errors from positioning computations is to perform many position computations, and to determine the average over all solutions. Many receivers allow the user to do this. It should, however, be clear from the above that *averaging* may address *random* errors such as signal noise, selective availability (SA) and multi-path to some extent, but not *systematic* sources of error, such as incorrect satellite data, atmospheric delays, and GDOP effects. These sources should be removed before averaging is applied. It has been shown that averaging over 60 min in absolute, single-point positioning based on code measurements, before systematic error removal, leads to only a 10–20% improvement of accuracy. In such cases, receiver

random and systematic error

averaging is therefore of limited value and requires near-optimal conditions for long periods. Averaging is a good technique if systematic errors have been accounted for.

In relative positioning, also known as *differential positioning*, one tries to remove some of the sources of systematic error by taking into account measurements of these errors in a nearby stationary *reference receiver* that has an accurately known position. By using these systematic error findings for the reference receiver, the position of the *target receiver* of interest can be determined much more precisely.

In an optimal setting, the reference and target receiver experience identical conditions and are connected by a direct data link, allowing the target to receive correctional data from the reference. In practice, relative positioning allows reference and target receiver to be 70–200 km apart; they will experience essentially similar atmospheric signal error. Selective availability can also be addressed in this away.

For each satellite in view, the reference receiver will determine its pseudorange error. After all, its position is known to a high degree of accuracy, so it can solve any pseudorange equations to determine the error. Subsequently, the target receiver, having received the error characteristics will apply the correction for each of the satellite signals that it uses for positioning. In doing so, it can improve its accuracy to within 0.5–1 m.

The discussion above assumes we needed positioning information in real time, which called for a data link between reference and target receiver. But various uses of satellite-based positioning do not need real time data, making post-processing of the recorded positioning data suitable. If the target receiver records time and position accurately, correctional data can be used later to improve the accuracy of the originally recorded data.

Finally, mention should be made of the notion of *inverted relative positioning*. The principles are still the same as above, but with this technique the target receiver does not correct for satellite pseudorange error, but rather uses a data link to upload its positioning/timing information to a central repository, where the corrections are applied. This can be useful in cases where many target receivers are needed and budget does not allow them to be expensive.

3.2.4 Network positioning

Now that the advantages of relative positioning have been discussed, we can move on to the notion of *network positioning*: an integrated, systematic network of reference receivers covering a large area, perhaps an entire continent or even the whole globe.

The organization of such a network can take different shapes, augmenting an already existing satellite-based system. Here we discuss a general architecture, consisting of a network of *reference stations*, strategically positioned in the area to be covered, each of them constantly monitoring signals and their errors for all positioning satellites in view. One or more *control centres* receive the reference station data, verify this for correctness, and relay (uplink) this information to a *geostationary satellite*. The satellite will retransmit any correctional data to the area that it covers, so that *target receivers*, using their own approximate position, can determine how to correct for satellite signal error, and consequently obtain much more accurate position fixes.

With network positioning, accuracy in the sub-metre range can be obtained. Typically, advanced receivers are required, but the technology lends itself also for solutions with a single advanced receiver that functions in the direct neighbourhood as a reference receiver to simple ones.

3.2.5 Code versus phase measurements

Up until this point, we have assumed that the receiver determines the range of a satellite by measuring time delay of the received ranging code. There exists a more advanced range determination technique, known as *carrier phase measurement*. This typically requires more advanced receiver technology and longer observation sessions. Currently, carrier phase measurement can only be used with relative positioning, as absolute positioning using this method is not yet well developed.

The technique aims to determine the number of cycles of the (sine-shaped) radio signal between sender and receiver. Each cycle corresponds to one wavelength of the signal, which in the L-band frequencies used is 19–24 cm. Since the number of cycles of the signal cannot be measured directly, it is determined (in a long observation session) from the change in carrier phase over time. Such a change occurs because the satellite is orbiting. From its orbit parameters and the change in phase over time, the number of cycles can be derived.

With relative positioning techniques, a horizontal accuracy of 2 mm–2 cm can be achieved. This degree of accuracy makes it possible to measure tectonic plate movements, which can be as large as 10 cm per year for some locations on the planet.

3.2.6 Positioning technology

This section provides information on currently available satellite-based positioning technology. At present, two satellite-based positioning systems are operational—GPS and GLONASS—and a third is in the implementation phase—Galileo. These systems are US, Russian and European, respectively. Any of them, but especially GPS and Galileo, will be improved over time and will be augmented with new techniques.

GPS

The NAVSTAR Global Positioning System (GPS) was declared operational in 1994, providing Precise Positioning Services (PPS) to US and allied military forces, as well as US government agencies; civilians throughout the world have access to Standard Positioning Services (SPS). The GPS space segment nominally consists of 24 satellites, each of which orbits our planet in 11 h 58 min at an altitude of 20,200 km. There can be any number of satellites active, typically between 21 and 27. The satellites are organized in six orbital planes, somewhat irregularly spaced, at an angle of inclination of 55–63° to the equatorial plane; nominally four satellites orbit in each plane (see Figure 3.28). This means that a receiver on Earth will have between five and eight (rarely, even up to 12) satellites in view at any moment in time. Software packages exist to help in planning GPS surveys, identifying the expected satellite set-up for any location and time.

The NAVSTAR satellites transmit two radio signals, an L1 frequency of 1575.42 MHz and an L2 frequency of 1227.60 MHz. There is also a third and fourth signal, but these are not important for the discussion here. The role of the L2 signal is to provide a second radio signal, thereby providing a way, with (more expensive) dual-frequency receivers, of determining fairly precisely the actual ionospheric delay of the satellite signals received.

GPS uses WGS84 as its reference system,which has been refined on several occasions and is now aligned with the ITRF at the level of a few centimetres worldwide. (See also Section 3.1.1.) GPS has adopted UTC as its time system.

WGS84

For civilian applications, GPS receivers of varying quality are available, their quality depending on the embedded positioning features: supporting single- or dual-frequencies; supporting only absolute or also relative positioning; performing code
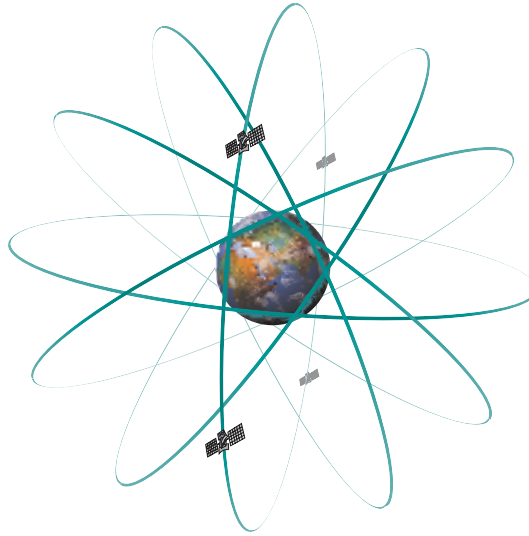
Figure 3.28
Constellation of satellites in the GPS system; here four satellites are shown in only one orbital plane.

measurements or also carrier phase measurements.

GLONASS

What GPS is to the US military, is GLONASS to the Russian military, specifically the Russian Space Forces. Both systems were primarily designed on the basis of military requirements, but GLONASS did not significantly develop civil applications as GPS did and thus it is commercially less important.

GLONASS's space segment consists nominally of 24 satellites, organized in three orbital planes, at an inclination of $64.8°$ to the Equator. Its orbiting altitude is 19,130 km, with a period of revolution of 11 h 16 min.

GLONASS uses the PZ-90 as its reference system and, like GPS, uses UTC as its time reference, albeit with an offset for Russian daylight.

GLONASS's radio signals are somewhat similar to those of GPS, differing only in the details: the frequency of GLONASS's L1 signal is approximately 1605 MHz (changes are underway), and its L2 signal approximately 1248 MHz; otherwise, GLONASS's system performance is rather comparable with that of GPS.

Galileo

In the 1990s, the European Union (EU) judged that it needed its own satellite-based positioning system, to become independent of the GPS monopoly and to support its own economic growth by providing services of high reliability under civilian control. The EU system is named Galileo.

The vision is that satellite-based positioning will become even bigger due to the emergence of mobile phones equipped with receivers, perhaps with some 400 million users by the year 2015. The development of the system has experienced substantial delays; currently European ministers insist that Galileo should be up and running by the end of 2013.

When completed, Galileo will have 27 satellites, with three in reserve, orbiting in one of three, equally spaced, circular orbits at an elevation of 23,222 km and inclined at $56°$ to the Equator. This higher inclination (when compared to that of GPS) has been chosen to provide better positioning coverage at high latitudes, such as in northern

Scandinavia, where GPS performs rather poorly.

In June 2004, the EU and the US agreed to make Galileo and GPS compatible by adopting interchangeable set-ups for their satellite signals. The effect of this agreement is that a Galileo/GPS tandem satellite system will have so many satellites in the sky (close to 60) that a receiver can almost always find an optimal constellation in view.

This will be especially useful in situations where in the past signal reception was poor, in built-up areas and forests, for instance. It will also bring the implementation of a Global Navigation Satellite System (GNSS) closer, since positional accuracy and reliability will improve. Such a system would bring the ultimate development of fully automated air and road traffic control systems much closer. Automatic aircraft landing, for instance, requires a horizontal accuracy in the order of 4 m, and a vertical accuracy of less than 1 m. Currently, these requirements cannot be reliably met.

The Galileo Terrestrial Reference Frame (GTRF) will be a realization of the ITRS and will be set up independently from that of GPS so that one system can back up the other. Positional differences between WGS84 and GTRF will be at worst only a few centimetres.

The Galileo System Time (GST) will closely follow International Atomic Time (TAI), with a time offset of less than 50 ns for 95% of the time over any period of the year. Information on the actual offset between GST and TAI, and between GST and UTC (as used in GPS), will be broadcast in the Galileo satellite signal.

Satellite-based augmentation systems

Satellite-based augmentation systems (SBAS) aim to improve the accuracy and reliability of satellite-based positioning (see Subsection 3.2.4) in support of safety-critical navigation applications, such as aircraft operations near airfields. Typically, these systems make use of an extra, now geostationary, satellite that has a large service area, for example a continent, and which sends differential data about standard positioning satellites that are currently in view in its service area. If multiple ground reference stations are used, the quality of the differential data can be quite good and reliable. Usually this satellite will use radio signals of the same frequency as those in use by the positioning satellites, so that receivers can receive the differential code without problem.

Not all advantages of satellite augmentation will be useful for all receivers. For consumer market receivers, the biggest advantage, as compared to standard relative positioning, is that SBAS provides an ionospheric correction grid for its service area, from which a correction specific for the location of the receiver can be retrieved. This is not true in relative positioning, where the reference station determines the error it experiences and simply broadcasts this information for nearby target receivers to use. With SBAS, the receiver obtains information that is best viewed as a geostatistical interpolation of errors from multiple reference stations.

More advanced receivers will be able to deploy also other differential data such as corrections on satellite position and satellite clock drift.

Currently, three systems are operational: for North America WAAS (Wide-Area Augmentation System) is in place; EGNOS (European Geostationary Navigation Overlay Service) for Europe; and MSAS (Multi-functional Satellite Augmentation System) for eastern Asia. The ground segment of WAAS consists of 24 control stations, spread over North America; that of EGNOS has 34 control stations. These three systems are compatible, guaranteeing international coverage.

Usually signals from the geostationary SBAS satellites (under various names, such as AOR, Artemis, IOR, Inmarsat, MTSAT) can be received even outside their respective

service areas. But the use of these signals there must be discouraged, as they will not help improve positional accuracy. Satellite identifiers, as shown by the receiver, have numbers above 30, setting them apart from standard positioning satellites.

Though originally intended to improve the safety of aircraft landings, SBAS, with its horizontal accuracy to within 3 m, has many other uses. At this level of accuracy, vehicle position can be determined to a specific road lane, and "automatic pilots" become a possibility.

Chapter 4

Sensors

Wim Bakker
Wan Bakx
Wietske Bijker
Karl Grabmaier
Lucas Janssen
John Horn
Gerrit Huurneman
Freek van der Meer
Christine Pohl
Klaus Tempfli
Valentyn Tolpekin
Tsehaie Woldai

4.1 Platforms and passive electro-optical sensors

Having explained the physics of sensing in Chapter 2, in this chapter we discuss sensor systems and set out to discover the logic of current electro-optical sensing technology. First, in Subsection 4.1.1, we will look at the characteristics of platforms used for geospatial data acquisition (GDA) from the air and from space: various platforms such as aircraft, space shuttles, space stations and satellites are used to carry one or more sensors for Earth Observation. Next, Subsection 4.1.2 will elaborate on frame and line cameras; the latter, which can be operated from the air or space, are also known as *pushbroom sensors*. Optical scanners (also referred to in the literature as across-track scanners or *whiskbroom scanners*) are treated in Section 4.1.3, which discusses multispectral, hyperspectral and thermal scanners in detail. Some camera systems can provide us with *stereo* images, justifying a short introduction to stereoscopy in Subsection 4.1.4.

4.1.1 Platforms and missions

Sensors used in Earth Observation can be operated at altitudes ranging from just a few centimetres above the ground—using field equipment—to those far beyond the atmosphere. Very often the sensor is mounted on a moving vehicle—which we call the *platform*—such as an aircraft or a satellite. Occasionally, static platforms are used. For example, we could mount a spectrometer on a pole to measure the changing re-

flectance of a specific crop during the day or over a whole season.

Moving platforms

To gain a wider view, we use aircraft at altitudes ranging up to approximately 20 km. Depending on the type of aerial survey and the weight of equipment and survey costs, we can choose from a variety of vehicles. Fixed-wing aircraft are used for thermal scanning and a systematic photo-coverage for topographic mapping, land titling projects, and the like. Aerial survey cameras are heavy and they are fixed to a stabilized mount set in a hole in the floor of the aircraft. Most survey airplanes fly lower than 8 km but higher than 1000 m. They can fly as slow as 150 km h^{-1}, but even at that speed image quality is already affected by motion blur unless the camera is fitted with a compensation device. Aerial survey cameras are highly sophisticated and expensive.

Airborne laser-scanner systems used to be heavy, but nowadays the total weight of the equipment can be as light as 30 kg. Laser scanners are either mounted on fixed-wing aircraft or helicopters, the latter being able to fly very slowly at low altitudes, thus allowing the acquisition of highly detailed data (at high costs per unit of area). The small-format cameras used are cheaper and lighter than large-format aerial survey cameras, making it possible to mount these systems on micro-light airplanes for urban reconnaissance, or even kites (e.g. for surveying an industrial area). Unmanned aerial vehicles (UAVs) are gaining popularity for observing dangerous areas or to reduce costs. A special type of UVA, the High Altitude Long Endurance (HALE) vehicle, can bridge the gap between manned survey aircraft and spacecraft or satellites. Typically, a HALE is a remotely operated aircraft of ultra-light weight and load that flies for months at altitudes of around 20 km.

A key advantage of aerial surveys is that they can be "targeted". The survey can be undertaken at exactly the required time and can be done with exactly the required spatial resolution by having the aircraft fly at the required altitude. Moreover, in comparison with civilian satellite RS, we can acquire images of much higher spatial resolution, enabling recognition of objects of much smaller size. With current aerial survey cameras, we can achieve a pixel size on the ground as small as 5 cm.

Satellites are launched by rocket into space, where they then circle the Earth for 5 to 12 years on a predefined orbit. The choice of orbit depends on the objectives of the sensor mission; orbit characteristics and different orbit types are explained below. A satellite must travel at high speed to orbit at a certain distance from the Earth; the closer to the Earth, the faster the speed required. A space station such as ISS has a mean orbital altitude of 400 km and travels at roughly 27,000 km h^{-1}. The Moon at a distance of 384,400 km can conveniently circle the Earth at only 3700 km h^{-1}. At altitudes of 200 km, satellites already encounter traces of the atmosphere, which causes rapid orbital and mechanical decay. The higher the altitude, the longer is the expected lifetime of the satellite. The majority of civilian Earth-observing satellites orbit at altitudes ranging from 500 to 1000 km. Here we generally find the "big boys", such as Landsat-7 (2200 kg) and Envisat (8200 kg), but the mini-satellites of the Disaster Management Constellation (DMC) also orbit in this range. DMC satellites have a weight of around 100 kg and were launched by several countries into space early in the current millennium at relatively low-cost. These satellites represent a network for disaster monitoring that provides images in three or four spectral bands with a ground pixel size of 32 m or smaller.

satellites

Satellites have the advantage over aerial survey of continuity. Meteosat-9, for example, delivers a new image of the same area every 15 minutes and it has done so every day for many years. The high temporal resolution at low cost goes together with a low

spatial resolution (pixel size on the ground of 1×1 km^2). Both the temporal and the spatial resolution of satellite remote sensors are fixed. While aerial surveys have been restricted in some countries, access to satellite RS data is commonly easier, although not every type of satellite RS image is universally available.

Aerial survey missions

Modern airborne sensor systems use a high-end GPS receiver and many also include an Inertial Measuring Unit (IMU). GPS is used for navigation and for coarse "sensor positioning". We need to know the coordinates of the exposure stations of a camera to relate points and features in the images to positions on the ground; differential GPS is applied for more precise positioning. To this end, we need a reference GPS station on the ground within some 30 km from the aircraft. Adding an IMU has two advantages: IMU readings can be used to improve the accuracy of the coordinates obtained by GPS (achieving a RMSE better than 0.1 m); and the IMU measures the attitude angles of the sensor (Figure 4.27). An IMU, an assemblage of gyros and accelerometers, is a sophisticated, heavy, and expensive instrument that was originally used only in Inertial Navigation Systems (INSs). Measuring continuously the position and attitude of the moving sensor, an IMU allows us to relate the sensor recordings to position in the terrain in near real-time. We call this *direct sensor orientation*. We need a GPS-IMU positioning and orientation system (POS) for line cameras and scanners; for frame cameras we can also solve the georeferencing problem indirectly (see Section 5.3).

direct sensor orientation
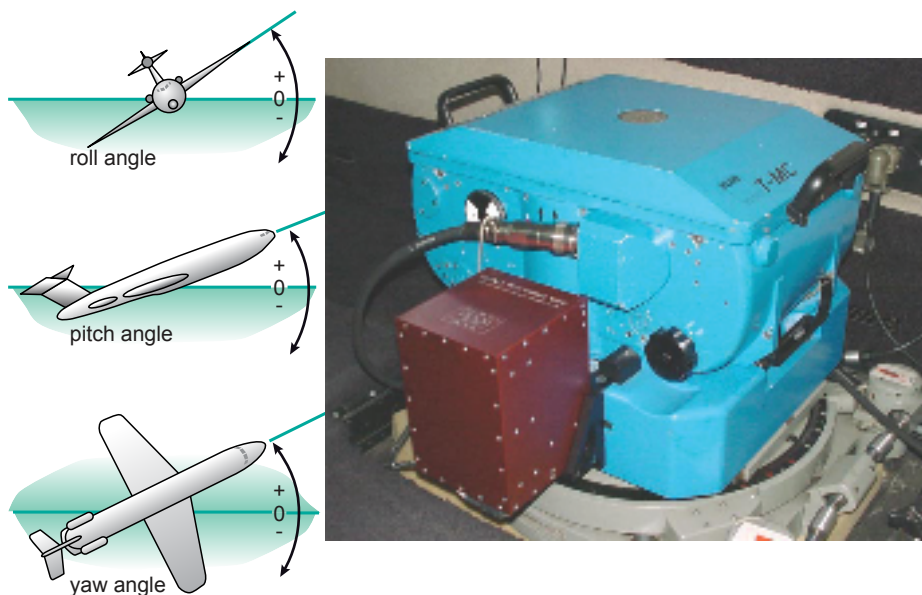
Figure 4.1
Attitude angles (left) and an IMU attached to a Zeiss RMK-TOP aerial camera (courtesy of IGI).

Mission planning and execution is usually done by commercial survey companies or, otherwise, by large national mapping agencies or the military. During missions, companies use professional software for flight planning and, most likely, one of the two integrated aircraft guidance and sensor management systems available (produced by APPLANIX or IGI). Pioneering work on computer-controlled navigation and camera management was done at ITC in the days when it still had an aerial photography and navigation department. The basics of planning aerial survey missions are explained in Section 4.6.

Satellite missions

The monitoring capabilities of a satellite-borne sensor are to a large extent determined by the parameters of the satellite's orbit. An *orbit* is a circular or elliptical path described by the satellite in its movement around the Earth. Different types of orbits are required to achieve continuous monitoring (meteorology), global mapping (land cover mapping) or selective imaging (urban areas). For Earth Observation, the following orbit characteristics are relevant:

- *Orbital altitude* is the distance (in km) from the satellite to the surface of the Earth. It influences to a large extent the area that can be viewed (i.e. the *spatial coverage*) and the details that can be observed (i.e. the *spatial resolution*). In general, the higher the altitude, the larger the spatial coverage but the lower the spatial resolution.

- *Orbital inclination angle* is the angle (in degrees) between the orbital plane and the equatorial plane. The inclination angle of the orbit determines, together with the field of view (FOV) of the sensor, the latitudes up to which the Earth can be observed. If the inclination is 60°, then the satellite orbits the Earth between the latitudes 60° N and 60° S. If the satellite is in a low-Earth orbit with an inclination of 60°, then it cannot observe parts of the Earth at latitudes above 60° North and below 60° South, which means it cannot be used for observations of the Earth's polar regions.

- *Orbital period* is the time (in minutes) required to complete one full orbit. For instance, if a polar satellite orbits at 806 km mean altitude, then it has an orbital period of 101 minutes. The Moon has an orbital period of 27.3 days. The speed of the platform has implications for the type of images that can be acquired. A camera on a low-Earth orbit satellite would need a very short exposure time to avoid motion blur resulting from the high speed. Short exposure times, however, require high intensities of incident radiation, which is a problem in space because of atmospheric absorption. It should be obvious that the contradictory demands of high spatial resolution, no motion blur, high temporal resolution, long satellite lifetime (thus lower cost) represent a serious challenge for satellite-sensor designers.

- *Repeat cycle* is the time (in days) between two successive identical orbits. The *revisit time* (i.e. the time between two subsequent images of the same area) is determined by the repeat cycle together with the pointing capability of the sensor. *Pointing capability* refers to the possibility of the sensor–platform combination to look to the side, or forward, or backward, and not only vertically downwards. Many modern satellites have such a capability. We can make use of the pointing capability to reduce the time between successive observations of the same area, to image an area that is not covered by clouds at that moment, and to produce stereo images (see Subsection 4.1.4).

The following orbit types are most common for remote sensing missions:

- *Polar orbit* refers to orbits with an inclination angle between 80° and 100°. An orbit having an inclination larger than 90° means that the satellite's motion is in a westward direction. Such a polar orbit enables observation of the whole globe, also near the poles. Satellites typically orbit at altitudes of 600–1000 km.

- *Sun-synchronous orbit* refers to a polar or near-polar orbit chosen in such a way that the satellite always passes overhead at the same time. Most Sun-synchronous

orbits cross the Equator mid-morning, at around 10:30 h local solar time. At that moment the Sun angle is low and the shadows that creates reveal terrain relief. In addition to day light images, a Sun-synchronous orbit also allows the satellite to record night images (thermal or radar, passive) during the ascending phase of the orbit on the night side of the Earth.

- A *Geostationary orbit* refers to orbits that position the satellite above the Equator (inclination angle: 0 °) at an altitude of approximately 36,000 km. At this distance, the orbital period of the satellite is equal to the rotational period of the Earth, exactly one sidereal day. The result is that the satellite has a fixed position relative to the Earth. Geostationary orbits are used for meteorological and telecommunication satellites.
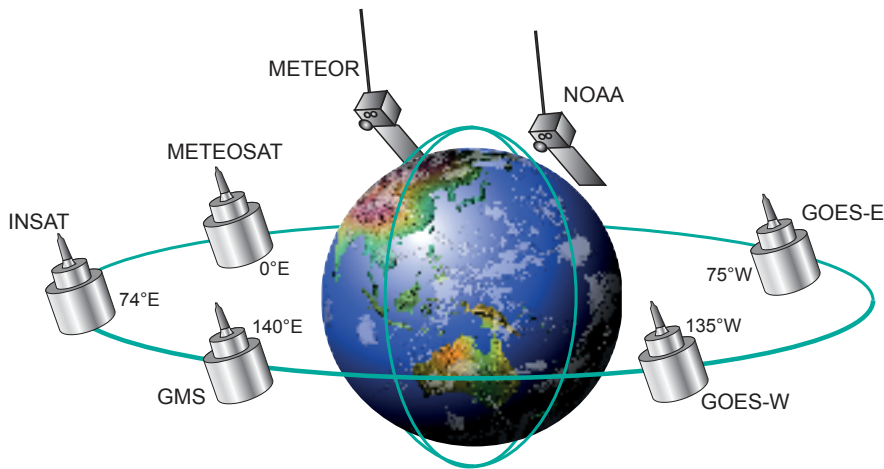
Figure 4.2
Meteorological observation by geostationary and polar satellites.

Today's meteorological weather satellite systems use a combination of geostationary satellites and polar orbiters (Figure 4.28). The geostationary satellites offer a continuous hemispherical view of almost half the Earth (45%), while the polar orbiters offer a higher spatial resolution.

RS images from satellites come with data on orbital parameters and other parameters to facilitate georeferencing of the images. High resolution sensor systems such as Ikonos or QuickBird use GPS receivers and star trackers as their POS.

The data from space-borne sensors need to be transmitted to the ground in some way. Russia's SPIN-2 satellite, with its KVR camera, used film cartridges that were dropped over a designated area on the Earth. Today's Earth Observing satellites *downlink* the data. The acquired data are sent directly to a receiving station on the ground, or via a geostationary communication satellite. One current trend is that small receiving units, consisting of a small dish with a PC, are being developed for local reception of RS data.

4.1.2 Cameras

A *digital camera* is an electro-optical remote sensor. In its simplest form, it consists of the camera body, a lens, a focal plane array of CCDs, and a storage device, but no mechanical component. The CCD array can either be an assembly of linear arrays or a matrix array (Figure 4.3). Accordingly, we talk about line cameras and frame cameras. A small-format frame camera has a single matrix chip and closely resembles a photographic camera. The chip (a) of the Figure 4.3 has three channels, one for each primary colour (red, green, blue); three elongated CCDs next to each other constitute

CCD

one square "colour pixel". Each CCD has its colour filter right on top to only transmit the required band of incident light. The linear chip (b) of the Figure 4.3 also has three channels; three lines of square CCDs are assembled next to each other. A line camera is exclusively used on a moving platform, which can be a car, an aircraft or a spacecraft. SPOT-1, launched in 1986, was the first satellite to use a line camera. Line cameras build up a digital image of an area line by line (Figure 4.4). In the older literature, therefore, it is also referred to as *pushbroom scanner*, as opposed to a *whiskbroom scanner* (see Subsection 4.1.3), which actually scans (across the track of the moving platform).
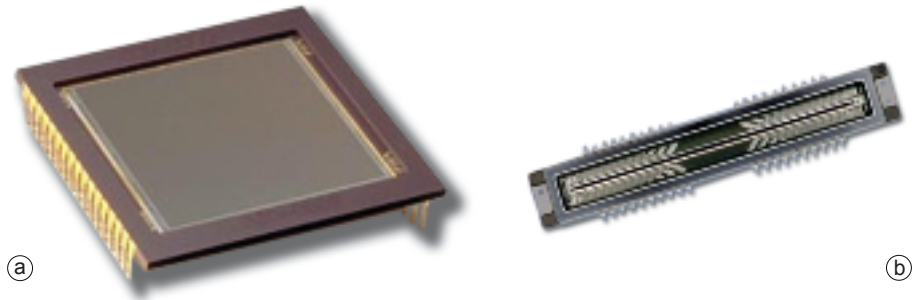
Figure 4.3
Two CCD chips: (a) matrix array Kodak KAF-16801, pixel size 9 μm; (b) linear array Kodak KLI-14403, pixel size 5 μm.

(a)   (b)
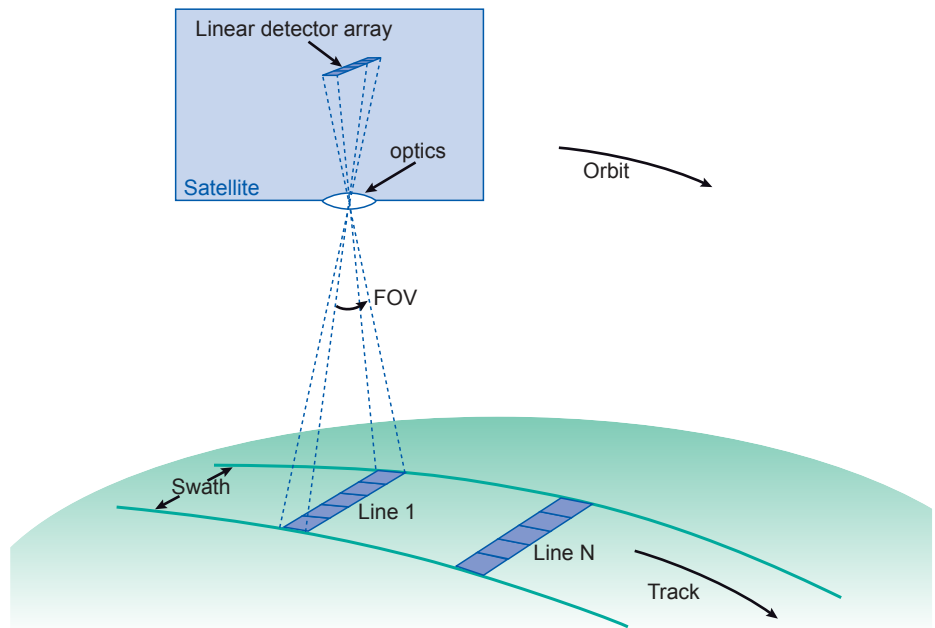
Figure 4.4
Principle of imaging by line camera on a spacecraft ('pushbrooming').

Detector arrays

Cameras are used for sensing in the visible, NIR, and SWIR portions of the spectrum. We need different types of semiconductors for sensing in this range; the semiconductors used are all solid-state detectors but are made of different material for different spectral ranges. CCDs are the most common type of semiconductor used today for sensing in the visible to very near-IR range; they are made of silicon.

The spectral sensitivity of a sensor band is commonly specified by a lower- and an upper-bound wavelength of the spectral band covered, e.g. 0.48 to 0.70 μm for the

SPOT-5 panchromatic channel. However, a detector such as a CCD is not equally sensitive to each monochromatic radiation within this band. The actual response of a CCD can be determined in the laboratory; an example of a resulting spectral response curve is shown in Figure 4.5. The lower and upper bound specification is usually chosen at the wavelengths where the 50% response is achieved. The DN produced by a detector results from averaging the spectral response of incident radiation. Figure 4.5 shows that the DNs of AVHRR channel 1 are biased towards red, whereas the brightness sensation of our eyes is dominated by yellow-green. The CCDs of a channel array do not have exactly the same sensitivity. It takes radiometric sensor calibration to determine the differences. CCDs show, moreover, varying degrees of degradation over time. Therefore, radiometric calibration needs to be done regularly. Knowing the detector's spectral sensitivity becomes relevant when we want to convert DNs to radiances (see Section 5.2).

spectral sensitivity
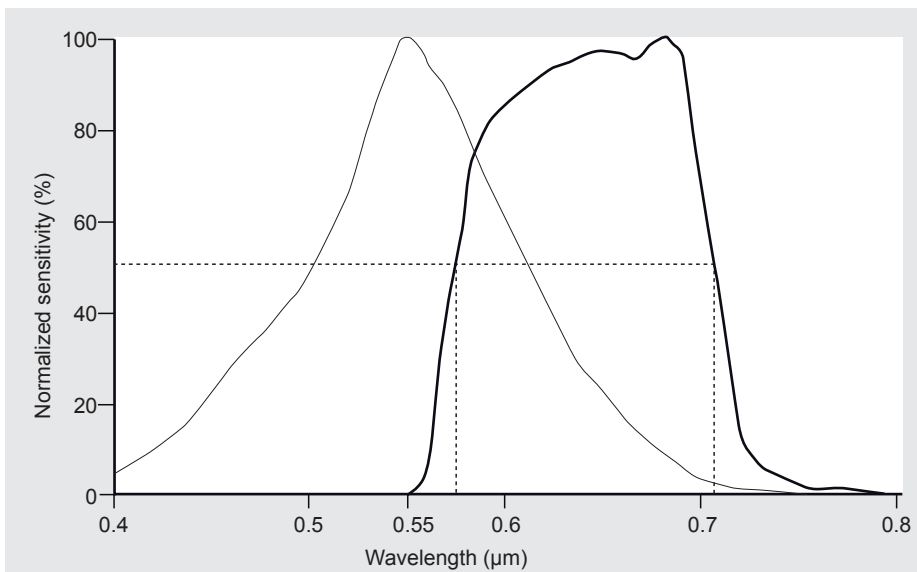
Figure 4.5
Normalized spectral response curve of (a) channel 1 of NOAA's AVHRR and (b) the spectral sensitivity of the rods of the human eye.

When compared with photographic film, most CCDs have a much higher general sensitivity and thus they need less light. The reason is that they typically respond to 70% of the incident light, whereas photographic film captures only about 2% of the incident light. They also offer a much better differentiation of intensity values in the very dark and the very bright parts of a scene.

If we were interested in a high radiometric resolution, we would like a CCD to have a wide dynamic range. *Dynamic range* is the ratio of the maximum to the minimum level of intensity that can be measured; it is also known as the signal to noise ratio of the detector. The maximum intensity is determined by the maximum charge capacity of the semiconductor cell. The minimum intensity is determined by the noise level. Noise is unwanted collected charge, for example caused by unblocked IR or UV radiation for a CCD that should be sensitive to blue light. It only makes sense to record a DN of many bits if the semiconductor cell has a wide dynamic range. It is the manufacturer's concern to ensure sufficient dynamic range to meet the radiometric resolution (expressed in bits) required by the user. We can compute the effective radiometric resolution of a sensor if the manufacturer specifies both the number of bits and the dynamic range.

dynamic range

We had line cameras in space and frame cameras in our pockets before we had any digital airborne camera offering satisfactory surveying quality. The main reason for

this is the ultra-high quality of aerial film cameras and their operational maturity, including the entire photogrammetric processing chain. Cameras on satellite platforms are exclusively line cameras, typically having a panchromatic channel and four more linear arrays (e.g. for red, green, blue, NIR). ASTER has two panchromatic channels, one linear array looking vertically down (nadir view) and the second looking backwards; the two resulting images can be used to generate stereo images. The first aerial line camera on the market was Leica's ADS40 (in 2000). It has three panchromatic detector arrays (forward, nadir, backward looking) and four multispectral ones (for RGB and NIR). One linear array consists of 12,000 CCDs.

Current CCD technology enables the production of very high quality linear arrays but not (yet) the very large matrix arrays that would be needed for large-format digital aerial cameras to be able to match the well-proven film-based survey camera. The two market leaders in digital aerial frame cameras, ZI and Microsoft (former Vexcel), therefore use several detector arrays for panchromatic imaging and software to compile a single large-format image from the sub-frames. ZI's DMC has, for example, $13,500 \times 7,500$ CCDs per sub-frame. One of the advantages of frame cameras is that the same photogrammetric software can be used as for photographs. At the moment there are about as many aerial line cameras as digital aerial frame cameras on the market.

Optical system

Cameras use either lenses or telescopes to focus incident radiation onto the focal plane where the CCD surfaces are. A lens of a simple hand-held camera is a piece of glass or plastic shaped to form an image by means of refraction. The lens cone of a survey camera contains a compound lens, which is a carefully designed and manufactured assembly of glass bodies (and thus very expensive). The camera head of a digital aerial frame camera (such as Z/I's DMC and Vexcel's UltraCam) even consists of several of such lenses to focus the light rays on the respective CCD arrays. However complicated a lens may be physically, geometrically imaging through a lens is simple. The geometric model that a point of an object connects to its point in the image by a straight line and that all such lines pass through the centre of the lens (Figure 4.6) is a very close approximation of reality. We refer to the geometric imaging of a camera as "central projection".

An important property of a lens is its *focal length*. The focal length, f, determines, together with the length of a CCD line, the FOV of the camera. The focal length together with the *flying height* determine the size of the ground-resolution cell for a given pixel size, p. The *flying height*, H, is either the altitude of the aircraft above the ground or the orbital altitude.

$$GRC = p\frac{H}{f} \tag{4.1}$$

The focal length of Leica's ADS40 line camera is 63 mm. At a flying height of 2000 m, we would attain a ground-resolution cell size in the across-track direction of 21 cm, provided the airplane flies perfectly horizontally over flat terrain (see Figure 4.6). You would conclude correctly that the ADS40 has a CCD/pixel size of 6.5 μm. The ratio $\frac{H}{f}$ is referred to as the *scale factor* of imaging.

Space-borne cameras do not have lens cones—they have telescopes. The telescopes of Ikonos and QuickBird consist of an assembly of concave and flat mirrors, thus achieving a spatial resolution that is absolutely amazing when considering their flying height. The focal length equivalent of the Ikonos telescope is 10 m. Ikonos specifications state a ground-resolution cell size of 80 cm for a panchromatic image at nadir.

linear arrays

matrix arrays

lens, focal length, scale

field of view

telescope

The size of a CCD determines the pixels size. A pixel projected onto the ground gives us the *ground resolution cell* (GRC) of the camera. The distance between the centres of two adjacent resolution cells of the same channel is called the *ground sampling distance* (GSD). Ideally the ground resolution cell size and the GSD are equal; the GSD then uniquely defines the spatial resolution of the sensor. This can be most easily achieved for panchromatic frame cameras. Note that the GSD is the same throughout an entire line if the terrain is flat and parallel to the focal plane (e.g. in the case of a nadir view of horizontal terrain); see Figure 4.6. If a space-borne line camera is pointed towards the left or the right of the orbit track (across-track, off-nadir viewing), we obtain an oblique image. The scale of an oblique image changes throughout the image. In the case of oblique viewing, Formula 4.1 does not apply anymore; the ground resolution cell size and the GSD increase with increasing distance from nadir. Section 5.3 explains how to deal with this.

ground resolution cell

ground sampling distance

Digital aerial cameras have several advantages over film cameras, pertaining to both the quality of images and economics. Digital aerial cameras commonly record in 5 spectral bands (panchromatic, RGB, NIR), therefore, we can obtain with one flight panchromatic stereo images, true colour images and false colour images; with a film camera we would have to fly this course three times and develop three different types of film. The radiometric quality of CCDs is better than that of photographic film. Digital cameras also allow an all-digital workflow, making processing faster and cheaper. Digital cameras can acquire images with a high likelihood of redundancy without additional costs for material and flying time; this favours automated information extraction. Finally, new geoinformation products can be generated as a result of various extended camera features. In Subsection 4.1.3 multispectral scanners are introduced. Line cameras as compared to across-track scanners have the advantage of better geometry. Airborne line cameras and scanners require gyroscopically stabilized mounts to reduce any effects of aircraft vibration and compensate for rapid movements of the aircraft. Such a stabilized mount keeps a camera in an accurate level position so that

advantages of digital cameras

it continuously points vertically downward. We want vertical images for mapping because of the better geometry. we, Therefore, also mount large-format digital frame cameras and film cameras on stabilized platforms for applications that require high-quality images.

4.1.3 Scanners

Components

An *optical scanner* is an electro-optical remote sensor with a scanning device, which is in most cases a mechanical component. In its simplest form (e.g. a thermal scanner operating in the 7 to 14 μm range), it consists of the sensor rack, a single detector with electronics, a mirror, optics for focusing, and a storage device (see Figure 4.7). A detector has a very narrow field of view (called the *instantaneous field of view* (IFOV)) of 2.5 milliradians or less. In order to image a large area, we have scan the ground across the track while the aircraft or space craft is moving. The most commonly-used scanning device is a moving mirror, which can be an oscillating mirror, a rotating mirror, or a nutating mirror. An alternative, which is used for laser scanning, is fiber optics.
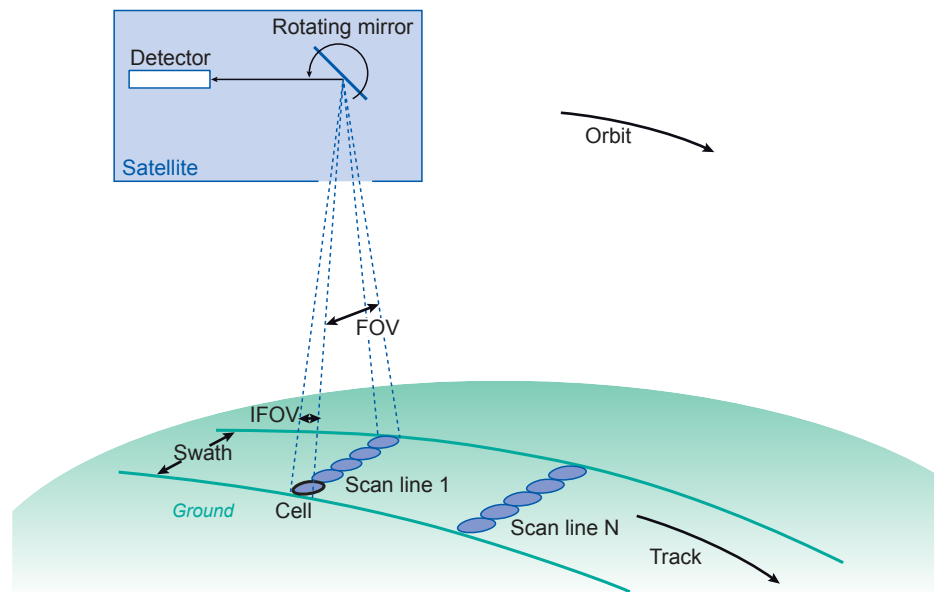
Figure 4.7
Principle of an across-track scanner.

detectors

Scanners are used for sensing in a broad spectral range, from light to TIR and beyond, to microwave radiation. Photodiodes made of silicon are used for the visible and NIR bands. Cooled photon detectors (e.g. using mercury-cadmium-telluride semiconductor material) are used for thermal scanners.

beam splitters

Most scanners are multispectral scanners, thus sensing in several bands, often including TIR (such as NOAA's AVHRR). As such, thermal scanners can be considered as being just a special type of multispectral scanner. A multispectral scanner has at least one detector per spectral band. Different from small-format frame cameras, for which filters are used to separate wavelength bands, scanners and line cameras use a prism and/or a grating as a *beam splitter*. A *grating* is a dispersion device used for splitting up SWIR and TIR radiation. Hyperspectral scanners also use gratings. A *prism* can split higher frequency radiation into red, green, blue, and NIR components. A simple

RGB and NIR scanner produces in one sweep of the mirror a single image line for each of the four channels.

Instead of using only one detector per band, space-borne scanners use several. The first civil space-borne remote sensor, Landsat MSS (launched in 1972), used six per band (thus, in total, 24; see Figure 2.18). ASTER uses 10 detectors for each of its five TIR channels. One sweep of the mirror of the ASTER thermal scanner produces, thus, 10 image lines for each of the five channels. If one channel should fail, only every 10th line of an image would be black. Section 5.2 treats the correcting of an image for periodic *line dropouts*.

Geometric aspects

At a particular instant, the detector of an across-track scanner observes an elliptical area on the ground, the ground resolution cell of the scanner. At nadir, the cell is circular, of diameter D. D depends on the IFOV, β, of the detector and the flying height.

$$D = \beta H \qquad (4.2)$$

A scanner with $\beta = 2.5$ mrad operated at $H = 4000$ m would have, therefore, a ground resolution of 10 m at nadir. Towards the edge of a swath, the ground resolution cell becomes elongated and bigger (Figure 4.29).
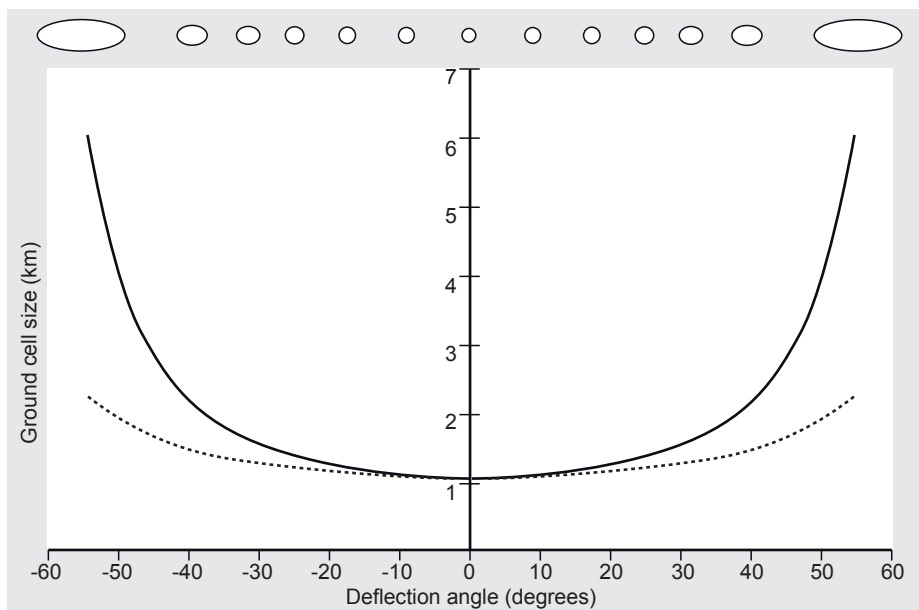
The width of the area that is covered by one sweep of the mirror, the *swath width*, depends on the FOV of the scanner. AVHRR has a very wide FOV of 110°; easy geometry was not a concern in the AVHRR design. Landast-7 has an FOV of only 15°, hence geometrically more homogeneous images result.

Reading out the detector is done at a fixed interval, the sampling interval. The sampling interval together with the speed of the moving mirror determines the GSD. The GSD can be smaller than D; we talk about *oversampling* if this is the case. The spatial resolution of the sensor is then not determined by the GSD but by the ground-resolution cell size (which is greater than or equal to D across the track).

4.1.4 Stereoscopy

Stereoscopy, the science of producing three-dimensional (3D) visual models using two-dimensional (2D) images, dates back to the 16th century. The astronomer Kepler was presumably the first person to define stereoscopic viewing. One of the main reasons for being able to perceive depth is that we have two eyes, which enables us to see a scene simultaneously from two viewpoints. The brain fuses the two 2D views into a three-dimensional impression. Judging which object is closer to us and which one is farther away with only one eye is only possible if we can use cues such as one object being partially obscured by the other one, or one appears smaller than the other although they are of the same size, etc. We can create the illusion of seeing three-dimensionally by taking two photographs or similar images and then displaying and viewing the pair simultaneously. Figure 4.30 illustrates the principle of stereoscopic viewing.

The advantage of stereoscopic viewing over monoscopic viewing (looking at a single image) is that image interpretation is easier, because we see the three-dimensional form of objects. Stereoscopy, moreover, has been the basis for 3D measurement by photogrammetry. Not just any two images can be viewed stereoscopically, they must fulfill several conditions. The same holds for making 3D measurements: we need at least two images and they must meet the preconditions. The basic requirements for a *stereo pair* are that the images of the same object or scene are taken from different positions, but not too far apart, and at a very similar scale. Different terms are used in stereoscopy, each with a slightly different meaning. A pair of images that meets the conditions of stereoscopic vision may be referred to as a stereo-image pair, a stereoscopic pair of images, stereo images, or simply as a stereo pair. A stereo pair arranged (on a computer monitor, on a table, or in a device) such that we can readily get a 3D visual impression may be called a stereograph, or stereogram). The 3D visual impression is called the stereo model, or stereoscopic model. We need special image-display techniques and stereoscopic viewing devices so that each eye sees only the image intended for it.
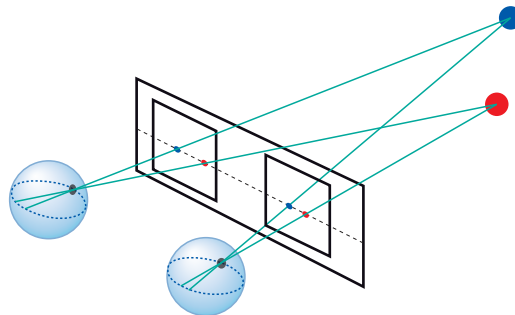
Figure 4.9
The principle of stereoscopy.

We have two options for obtaining a stereo pair with a space-borne sensor: a) use across-track pointing to image the same area from two different tracks, or b) apply along-track forward or backward viewing in addition to nadir viewing. The advantage of *in-track stereo* is that the two images are radiometrically very similar, because they are taken either at the same time or in quick succession; hence season, weather, scene illumination, and plant status are the same. In order to obtain a systematic coverage of an area with stereo images using an airborne frame camera, we need to take strips of vertical photos/images such that the images overlap by at least 60% (see Section 4.6).

4.2 Thermal remote sensing

Thermal remote sensing is based on the measuring of electromagnetic radiation in the infrared region of the spectrum. The wavelengths most commonly used are those in the intervals 3–5 μm and 8–14 μm, in which the atmosphere is fairly transparent and the signal is only slightly attenuated by atmospheric absorption. Since the source of the radiation is the heat of the imaged surface itself (see Figures 2.6 and 2.16), the handling and processing of TIR data is considerably different from remote sensing based on reflected sunlight:

- The surface temperature is the main factor that determines the amount of emitted radiation measured in the thermal wavelengths. The temperature of an object varies greatly depending on time of day, season, location, exposure to solar irradiation, etc. and is difficult to predict. In reflectance remote sensing, on the other hand, the incoming radiation from the Sun is considered constant and can be readily calculated, although atmospheric correction has to be taken into account.

- In reflectance remote sensing, the characteristic property we are interested in is the *reflectance* of the surface at different wavelengths. In thermal remote sensing, however, the one property we are interested in is, rather, how well radiation is *emitted* from the surface at different wavelengths.

- Since thermal remote sensing does not depend on reflected sunlight, it can also be done at night (for some applications this is even better than during the day).

4.2.1 Radiant and kinetic temperatures

The actual measurements by a TIR sensor will relate to the "spectral radiance" (measured in W m^{-2} sr^{-1} μm^{-1}) that reaches the sensor for a certain wavelength band. We know that the amount of radiation from an object depends on its temperature T and emissivity ϵ. That means that a cold object with high emissivity can radiate just as much radiation as a considerably hotter one with low emissivity. Often the emissivity of the object is unknown. If we assume that the emissivity of the object is equal to 1.0, then with the help of Planck's law we can calculate directly the ground temperature that is needed to create this amount of radiance in the specified wavelength band of the sensor for the object with a perfect emissivity. The temperature calculated in this way is the *radiant temperature* or T_{rad}. The terms *brightness* or "top-of-the-atmosphere" temperature are also frequently used.

radiant temperature

The radiant temperature calculated from the emitted radiation is in most cases lower than the true, *kinetic temperature* (T_{kin}) that we could measure on the ground with a contact thermometer. The reason for this is that most objects have an emissivity lower than 1.0 and radiate incompletely. To calculate the true T_{kin} from the T_{rad}, we need to know or estimate the emissivity. The relationship between T_{kin} and T_{rad} is:

kinetic temperature

$$T_{rad} = \epsilon^{1/4} T_{kin}. \tag{4.3}$$

With a single thermal band (e.g. Landsat-7 ETM+), ϵ has to be estimated from other sources. One way of doing this is to do a land cover classification with all available bands and then assign an ϵ value for each class from an emissivity table (e.g. 0.99 for water, 0.85 for granite).

In multispectral TIR, several bands of thermal wavelengths are available. With emissivity in each band, as well as the surface temperature (T_{kin}), unknown, we still have

an under-determined system of equations. For this reason, it is necessary to make certain assumptions about the shape of the emissivity spectrum we are trying to observe. Different algorithms exist to separate the influence of temperature from the emissivity.

4.2.2   Thermal applications

In general, applications of thermal remote sensing can be divided into two groups. In one group, the main interest is the study of surface composition by observing the surface emissivity in one or more wavelengths. In the other group, the focus is on surface temperature and its spatial and temporal distribution. The following discussion only concerns this second group.

Thermal hotspot detection   Another application of thermal remote sensing is the detection and monitoring of small areas with thermal anomalies. The anomalies can be related to fires, such as forest fires or underground coal fires, or to volcanic activity, such as lava flows and geothermal fields. Figure 4.10 shows an ASTER scene that was acquired at night. The advantage of night images is that the Sun does not heat up the rocks surrounding the anomaly, as would be the case during the day. This results in higher contrast between the temperatures of the anomaly itself and surrounding rocks. This particular image was acquired over the Wuda coal-mining area in China in September 2002. Hotter temperatures are represented by brighter shades of grey. On the right side, the Yellow River is clearly visible, since water does not cool down as quickly as the land surface does, due to thermal inertia. Inside the mining area (white box in Figure 4.10), several hotspots, with higher temperatures compared to the surrounding rocks, are visible. The inset shows the same mining area slightly enlarged. The hottest pixels are orange and show the locations of coal fires. If images are taken several weeks, or even years, apart the development of these underground coal fires, as well as the effect of fire fighting efforts, can be monitored quite effectively with thermal remote sensing.

Glaciers monitoring   With thermal remote sensing, studies of glaciers can go further than the plain observation of their extent. Understanding the dynamics of a glacier's state requires environmental variables. Ground surface temperature is obviously among the most important variables that affect glacier dynamics.

Urban heat islands   The temperature of many urban areas is significantly higher than that of surrounding natural and rural areas. This phenomenon is referred to as an urban heat island. The temperature difference is usually larger at night than during the day and occurs mainly due to the change of matter covering the land as a result of urban development: land cover in built-up areas retains heat much better than land cover in natural and rural areas. This affects the environment in many ways: it modifies rainfall patterns, wind patterns, air quality, the seasonality of vegetation growth, and so on. Urban heat islands also affect the health of urban inhabitants: in particular, they can modify the duration and magnitude of heat waves in urban areas, leading to increases in mortality rates. There are several ways to mitigate the urban heat island effect, the most prominent ones being the use of highly reflective materials and increasing the amount of urban vegetation. To study the urban heat island effect we need to observe the temperature in urban and surrounding areas. Thermal remote sensing is a suitable tool as it provides temperature measurements that incorporate the spatial extent of cities and their surroundings.
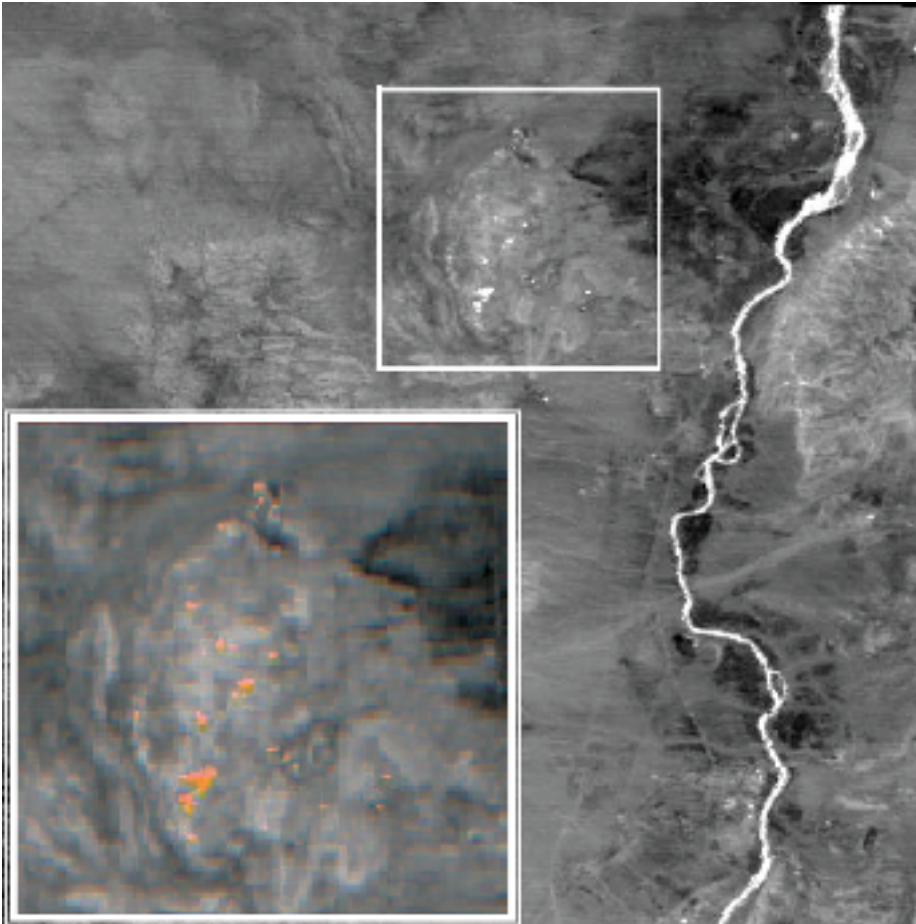
Figure 4.10
ASTER thermal band 10 over
Wuda, China. Light coloured
pixels inside the mining area
(white box) are caused mainly
by coal fires. Inset: pixels
exceeding the background
temperature of 18 °C are
orange for better visibility of
the fire locations. This scene
is approximately 45 km wide.

4.3   Imaging Spectrometry

You have learnt in Section 2.5 that materials of interest may be distinguished by their spectral reflectance curves (e.g. Figure 2.14). In this section we will call spectral reflectance curves *reflectance spectra*. Most multispectral sensors that were discussed in Chapter 2 acquire data in a number of relatively broad wavelength bands. However, typical diagnostic absorption features, characterizing materials of interest in reflectance spectra, are in the order of 20–40 nm in width. Hence, broadband sensors under-sample this information and do not allow full exploitation of the spectral resolution potential available. Imaging spectrometers typically acquire images in a large number of spectral bands (more than 100). These bands are narrow (less than 10–20 nm in width) and contiguous (i.e. adjacent), which enables the extraction of reflectance spectra at pixel scale (Figure 4.11). Such narrow spectra enable the detection of diagnostic absorption features. Different names have been coined for this field of remote sensing, including imaging spectrometry, imaging spectroscopy and hyperspectral imaging.

Figure 4.12 illustrates the effect of spectral resolution for the mineral kaolinite. From top to bottom, the spectral resolution increases from 100–200 nm (Landsat), 20–30 nm (GERIS), 20 nm (HIRIS), 10 nm (AVIRIS), to 1–2 nm (USGS laboratory reference spec-

trum). With each improvement in spectral resolution, the diagnostic absorption features and, therefore, the unique shape of kaolinite's spectrum become more apparent.
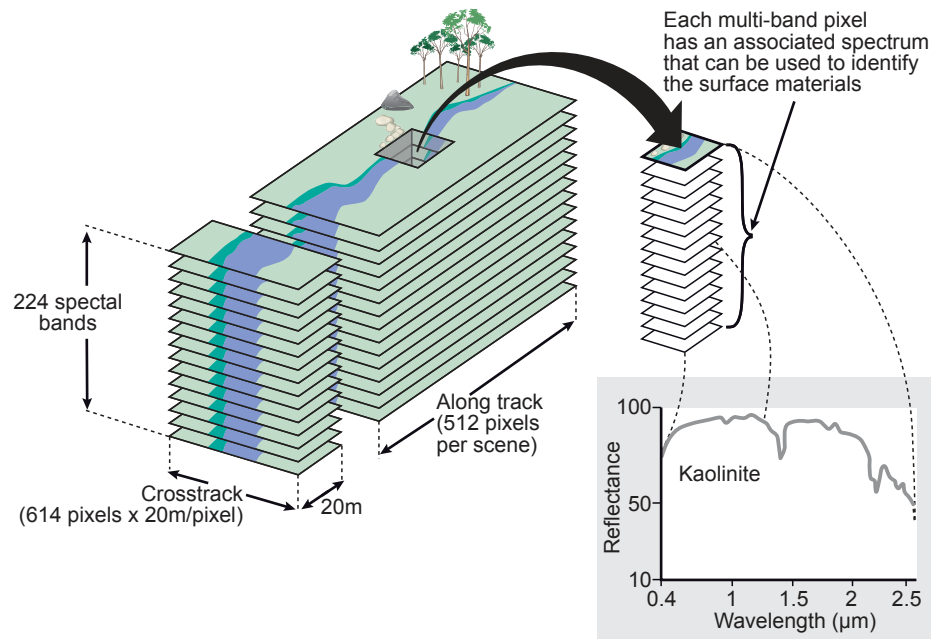
Figure 4.11
The concept of imaging spectrometry (adapted from [114]).

4.3.1 Reflection characteristics of rocks and minerals

reflectance spectra

Rocks and minerals reflect and absorb electromagnetic radiation as a function of the wavelength of the radiation. Reflectance spectra show these variations in reflection and absorption for various wavelengths (Figure 4.13). By studying the reflectance spectra of rocks, individual minerals and groups of minerals may be identified. In the Earth sciences, absorption in the wavelength region 0.4 μm–2.5 μm is commonly used to determine the mineralogical content of rocks. In this region, various groups of minerals have characteristic reflectance spectra; examples include phyllosilicates, carbonates, sulphates, and iron oxides and iron hydroxides. High-resolution reflectance spectra for mineralogy studies can easily be obtained in the field or the laboratory using field spectrometers.

Processes that cause absorption of electromagnetic radiation occur at the molecular and atomic levels. Two types of processes are important in the 0.4 μm–2.5 μm range: electronic processes; and vibrational processes ([21]). Depending on the molecular structure and composition, different absorption features can be identified. Reflectance spectra also correspond closely to the crystal structure of minerals and can, therefore, be used to obtain information about their crystallinity and chemical composition.

4.3.2 Pre-processing of imaging spectrometer data

radiometric calibration

atmospheric correction

Pre-processing of imaging spectrometer data involves radiometric calibration (see Section 5.2), which provides transfer functions to convert DN values to at-sensor radiance. The at-sensor radiance data have to be corrected by the user for atmospheric effects to obtain at-sensor or surface reflectance data. Section 5.2 contains an overview of the use of radiative transfer models for atmospheric correction. The correction provides absolute reflectance data, because the atmospheric influence is modelled and removed.

Figure 4.12
Example of a kaolinite spectrum at the original resolution (source: USGS laboratory) and at the spectral resolutions of various imaging devices. Note that the spectra are progressively offset upwards by 0.4 units for clarity (adapted from USGS).

Alternatively, users can make a scene-dependent relative atmospheric correction using empirically derived models for the radiance-reflectance conversion that are based on calibration targets found in the imaging spectrometer data set. Empirical models often used include techniques known as flat-field correction and empirical-line correction. Flat-field correction achieves radiance-reflectance conversion by dividing the whole data set on a pixel-by-pixel basis by the mean value of a target area within the scene that is spectrally and morphologically flat, spectrally homogeneous and has a high albedo. Conversion of raw imaging spectrometer data to reflectance data using the empirical-line method, on the other hand, requires selection and spectral characterization (in the field with a spectrometer) of two calibration targets (a dark and a bright target). This empirical correction uses a constant gain and offset for each band to force a best fit between sets of field and image spectra that characterize the same ground areas, thus removing atmospheric effects, residual instrument artefacts, and viewing geometry effects.

relative correction

4.3.3 Applications of imaging spectrometry data

A brief outline of current applications in various fields relevant to the thematic context of ITC are described in the remainder of this subsection.

Geology and resources exploration

Imaging spectrometry is used by the mining industry for surface mineralogy mapping, to aid in ore exploration. Other applications of this technology include lithological and structural mapping. The petroleum industry is also developing methods for using imaging spectrometry for reconnaissance surveys. The main targets are hydrocarbon seeps and microseeps.

Figure 4.13
Effects of electronic and vibrational processes on absorption of electromagnetic radiation.
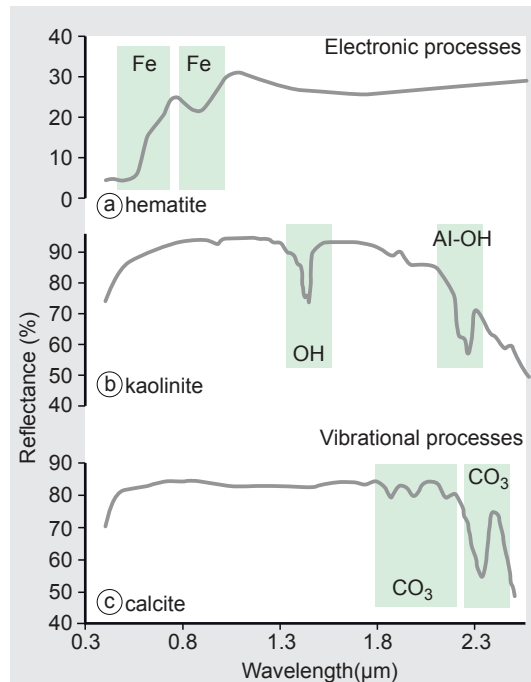
Other fields of application include environmental geology (and related geo-botany), in which currently much work is being done on acid mine drainage and mine-waste monitoring. Imaging of the atmospheric effects resulting from geological processes (e.g. sulfates emitted from volcanoes), to predict and quantify the presence of various gases for hazard assessment, is also an important field. In soil science, much emphasis has been placed on the use of spectrometry for the study of soil surface properties and soil composition analysis. Major elements such as iron and calcium, in addition to cation–anion exchange capacity, can be estimated from imaging spectrometry. In a more regional context, imaging spectrometry has been used to monitor agricultural areas (per-lot monitoring) and semi-nature areas. Recently, spectral identification from imaging spectrometers has been successfully applied to the mapping of the swelling clay minerals smectite, illite and kaolinite, in order to quantify the swelling potential of expansive soils. It should be noted that mining companies and, to a lesser extent, petroleum companies are already using imaging spectrometer data for reconnaissance-level exploration.

Vegetation sciences

Much research in vegetation studies has emphasized leaf biochemistry and leaf and canopy structure. Biophysical models for leaf constituents are currently available, as are soil–vegetation models. Estimates of plant material and structure, and biophysical variables, include carbon balance, yield/volume, nitrogen, cellulose, and chlorophyll. Leaf area index and vegetation indices have been extended to the hyperspectral domain and remain important physical variables for characterizing vegetation. One ultimate goal is the estimation of biomass and the monitoring of changes therein. Several research groups have been investigating the bi-directional reflectance function in relation to vegetation species analysis and floristics. Vegetation stress as a result of water deficiency, pollution (such as acid mine drainage) and geo-botanical anomalies in relation to ore deposits or petroleum and gas seepage links vegetation analysis to

exploration. Another upcoming field of application is precision agriculture, in which imaging spectrometry is being used to improve agricultural practices. An important factor in the health of vegetation is chlorophyll absorption and, in relation to that, the position of the red edge, determined using the red-edge index. Red edge is the name given to the steep increase in the reflectance spectrum of vegetation between visible red and near infrared wavelengths.

Hydrology

In hydrological sciences, the interaction of electromagnetic radiation with water, and the inherent and apparent optical properties of water are a central issue. Atmospheric correction and air–water interface corrections are very important in the imaging spectrometry of water bodies. Water quality of freshwater aquatic environments, estuarine environments and coastal zones usually has an important impact on national water bodies. Detection and identification of phytoplankton biomass, suspended sediments and other matter, coloured dissolved organic matter, and aquatic vegetation (i.e. macrophytes) are crucial variables in optical models of water quality. Much emphasis has been put on the mapping and monitoring of the state and growth or breaking down of coral reefs, as these are important for the CO_2 cycle. In general, many multi-sensor missions such as Terra and Envisat are directed towards integrated approaches for global climate change studies and global oceanography. Atmospheric models are important in global climate-change studies and aid in the correctin of optical data for scattering and absorption owing to trace gases in the atmosphere. In particular, the optical properties and absorption characteristics of ozone, oxygen, water vapour and other trace gases, and scattering by molecules and aerosols, are important variables in atmosphere studies. All these can be and are estimated from imaging spectrometry data.
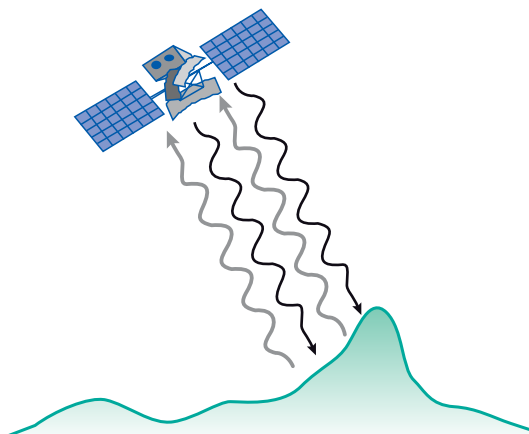
Figure 4.14
Principle of active microwave remote sensing.

4.4 Radar

4.4.1 What is radar?

Microwave remote sensing uses electromagnetic waves with wavelengths between 1 cm and 1 m (Figure 2.5). These relatively long wavelengths have the advantage that they can penetrate clouds and are not affected by atmospheric scattering. Although microwave remote sensing is primarily considered to be an active technique, passive sensors are also used. Microwave radiometers operate, similarly to thermal sensors, by detecting naturally emitted microwave radiation (either terrestrial or atmospheric). They are primarily used in meteorology, hydrology and oceanography.

microwave RS

In active systems, the antenna emits microwave signals to the Earth's surface, where they are backscattered. The part of the electromagnetic radiation that is scattered back in the direction of the antenna is detected by a sensor, as illustrated in Figure 4.14. There are several advantages to be gained from using active sensors, which have their own source of EM radiation:

- it is possible to acquire data at any time, also at night (similar to thermal remote sensing);

- since the waves are created by the sensor itself, the signal characteristics are fully controlled (wavelength, polarization, incidence angle, etc.) and can therefore be adjusted according to the desired application.

Active sensors can be divided into two types: imaging and non-imaging sensors. Radar sensors are typically active imaging microwave sensors. The term *radar* is an acronym for radio detection and ranging. *Radio* stands for the microwave component and *ranging* is another term for distance. Radar sensors were originally developed and used by the military. Nowadays, radar sensors are also widely used in civilian applications, such as environmental monitoring. Examples of non-imaging microwave instruments are *altimeters*, which collect distance information (e.g. sea-surface elevation), and *scatterometers*, which acquire information about object properties (e.g. wind speed).

non-imaging radar

The following subsection focuses on the principles of imaging radar and its applications. The interpretation of radar images is less intuitive than the interpretation of photographs and similar images. This is because of differences in the physical inter-

action of the waves with the Earth's surface. The interactions that take place and how radar images can be interpreted are also explained.

4.4.2 Principles of imaging radar

Imaging radar systems have a number of components: a transmitter, a receiver, an antenna, and a recorder. The transmitter is used to generate the microwave signal and transmit the energy to the antenna, from where it is emitted towards the Earth's surface. The receiver accepts the backscattered signal reaching the antenna and filters and amplifies it as required for recording. The recorder then stores the received signal.

Imaging radar acquires an image in which each pixel contains a digital number according to the strength of the backscattered radiation received from the ground. The radiation received from each emitted radar pulse can be expressed in terms of the physical variables and illumination geometry according to the *radar equation*:

backscattered radiation

$$P_r = \frac{G^2 \lambda^2 P_t \sigma}{(4\pi)^3 R^4},$$

(4.4)

where

P_r	=	received radiance,
G	=	antenna gain,
λ	=	wavelength,
P_t	=	emitted radiance,
σ	=	*radar cross-section*, which is a function of the object characteristics and the size of the illuminated area, and
R	=	range from the sensor to the object.

This equation demonstrates that there are three main factors that influence the strength of the backscattered radiation received:

- radar system properties, i.e. wavelength, antenna and emitted power;

- radar imaging geometry, which defines the size of the illuminated area, which is in turn a function of, for example, beam width, incidence angle and range;

- the characteristics of interaction of the radar signal with objects, i.e. surface roughness and composition, and terrain relief (magnitude and orientation of slopes).

  These factors are explained in more detail below.

What exactly does a radar system measure? To interpret radar images correctly, it is important to understand what a radar sensor detects. Radar waves have the same physical properties as those explained in Chapter 2. Radar waves, too, have electric and magnetic fields that oscillate as a sine wave in perpendicular planes. In dealing with radar, the concepts of wavelength, period, frequency, amplitude, and phase are therefore relevant.

A radar transmitter creates microwave signals, i.e. *pulses* of microwaves at a fixed frequency (the *Pulse Repetition Frequency*), that are directed by the antenna into a beam. A pulse travels in this beam through the atmosphere, "illuminates" a portion of the Earth's surface, is backscattered and passes through the atmosphere back to the antenna, where the signal is received and its intensity measured. The signal needs to travel twice the distance between an object and the receiver/antenna. As we know

intensity
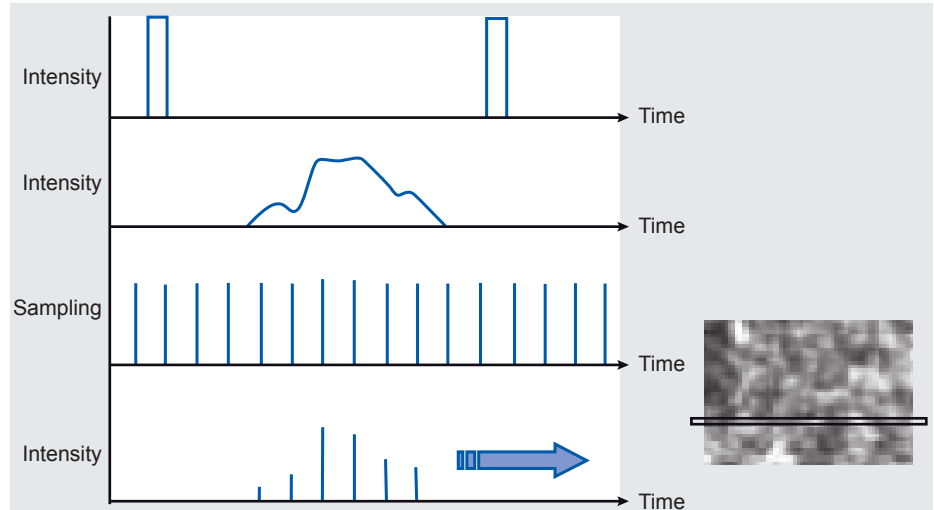
Figure 4.15
Illustration of how radar pixels result from pulses. For each sequence shown, one image line is generated.

the speed of light, we can calculate the distance (*range*) between sensor and object (see Formula 4.5).

To create an *image*, the return signal of each single pulse is sampled and samples stored in an image line (Figure 4.15). With the movement of the sensor while emitting pulses, a two-dimensional image is created (each pulse defines one line). The radar sensor therefore measures distances and backscattered signal intensities.

Commonly-used imaging radar bands   Similarly to optical remote sensing, radar sensors operate within one or more different bands. For better identification, a standard has been established that defines various wavelength ranges using letters to distinguish them from each other (Figure 4.16); you can recognize the different wavelengths used in radar missions from the letters used. The European ERS mission and the Canadian Radarsat use, for example, C-band radar. Just like multispectral bands, different radar bands provide information about different object characteristics.

Figure 4.16
The microwave spectrum and band identification by letters.

Band	**P**	**L**	**S**	**C**	**X**	**K**	**Q**	**V**	**W**
Frequency (GHz)	0.3	1.0	3.0		10.0	30.0			100.0
Wavelength (cm)	100	30	10		3	1			0.3

Microwave polarizations   The polarization of an electromagnetic wave is important in radar remote sensing. Depending on the orientation of the emitted and received radar wave, polarization will result in different images (see Figure 2.1, which shows a vertically polarized EM wave). It is possible to work with horizontally-, vertically- or cross-polarized radar waves. Using different polarizations and wavelengths, you can collect information that is useful for particular applications, e.g. to classify agricultural fields. In radar system descriptions you will come across the following abbreviations:

- HH: horizontal transmission and horizontal reception;

- VV: vertical transmission and vertical reception;
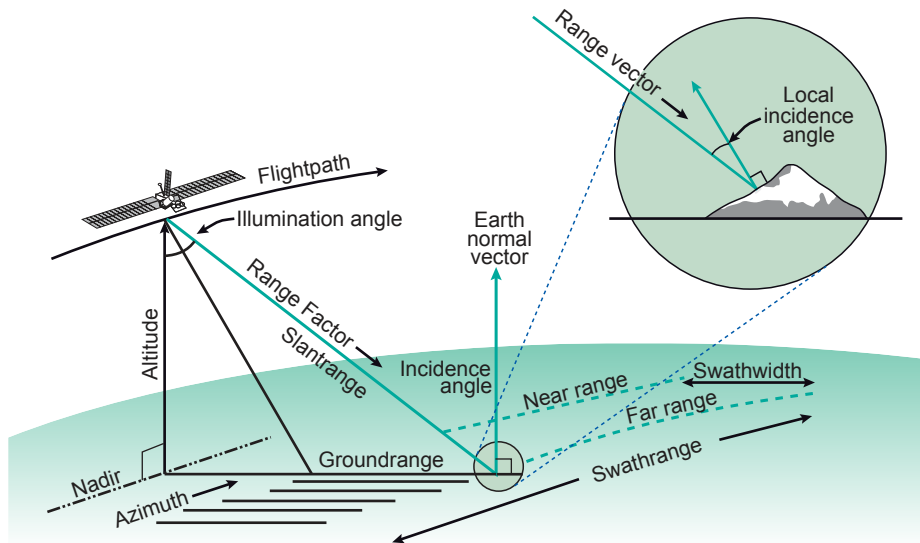
Figure 4.17
Radar remote-sensing
geometry.

- HV: horizontal transmission and vertical reception;

- VH: vertical transmission and horizontal reception.

4.4.3 Geometric properties of radar

The platform carrying the radar sensor travels along its orbit or flight path (Figure 4.17). You can see the ground track of the orbit/flight path on the Earth's surface at nadir. The microwave beam illuminates an area, or *swath*, on the Earth's surface, with an offset from nadir, i.e. side-looking. The direction along-track is called *azimuth* and the direction perpendicular (across-track) is called *range*.

azimuth

Radar viewing geometry

Radar sensors are side-looking instruments. The portion of the image that is closest to the nadir track of the satellite carrying the radar is called *near range*. The part of the image that is farthest from nadir is called *far range* (Figure 4.17). The *incidence angle* of the system is defined as the angle between the radar beam and the local Earth normal vector. Moving from near range to far range, the incidence angle increases. It is important to distinguish between the incidence angle of the sensor and the *local incidence angle*, which differs depending on terrain slope and the curvature of the Earth (Figure 4.17). The local incidence angle is defined as the angle between the radar beam and the local surface normal vector. The radar sensor measures the distance between antenna and object. This line is called the *slant range*. The true horizontal distance along the ground corresponding to each point of measured slant range is called the *ground range* (Figure 4.17).

ranges

Spatial resolution

In radar remote sensing, the images are created from the backscattered portion of emitted signals. Without further sophisticated processing, the spatial resolutions of slant range and azimuth direction are defined by the pulse length and the antenna beam width, respectively. This setup is called *real aperture radar* (RAR). As different parameters determine the spatial resolution in range and azimuth, it is obvious that the spatial

RAR

resolution in each direction is different from the other. For radar image processing and interpretation it is useful to resample the data to the same GSD in both directions.

Slant range resolution   For slant range, the spatial resolution is defined as the distance that two objects on the ground have to be apart to give two different echoes in the return signal. Two objects can be resolved in range direction if they are separated by at least half a pulse length. In that case, the return signals will not overlap. Slant range resolution is independent of the actual range (see Figure 4.18).

Azimuth resolution   The spatial resolution in azimuth direction depends on the beam width and the actual range. The radar beam width is proportional to the wavelength and inversely proportional to the antenna length, i.e. *aperture*. This means the longer the antenna, the narrower the beam and the higher the spatial resolution in azimuth direction.

aperture

RAR systems have their limitations in getting useful spatial resolutions of images because there is a physical limit to the length of the antenna that can be carried on an aircraft or satellite. On the other hand, shortening the wavelength will reduce the capability of penetrating clouds. To improve the spatial resolution, a large antenna is synthesized by taking advantage of the forward motion of the platform. Using all the backscattered signals in which a contribution of the same object is present, a very long antenna can be synthesized. This length is equal to the part of the orbit or flight path in which the object is "visible". Most airborne and space-borne radar systems use this type of radar. Systems using this approach are referred to as *Synthetic Aperture Radar (SAR)*.

SAR

4.4.4   Distortions in radar images

Due to the side-looking geometry, radar images suffer from serious geometric and radiometric distortions. In a radar image, you encounter variations in scale (caused by slant range to ground range conversion), *foreshortening*, *layover* and *shadows* (due to terrain elevation; see Figure 4.19). Interference due to the coherence of the signal causes *speckle* effects.
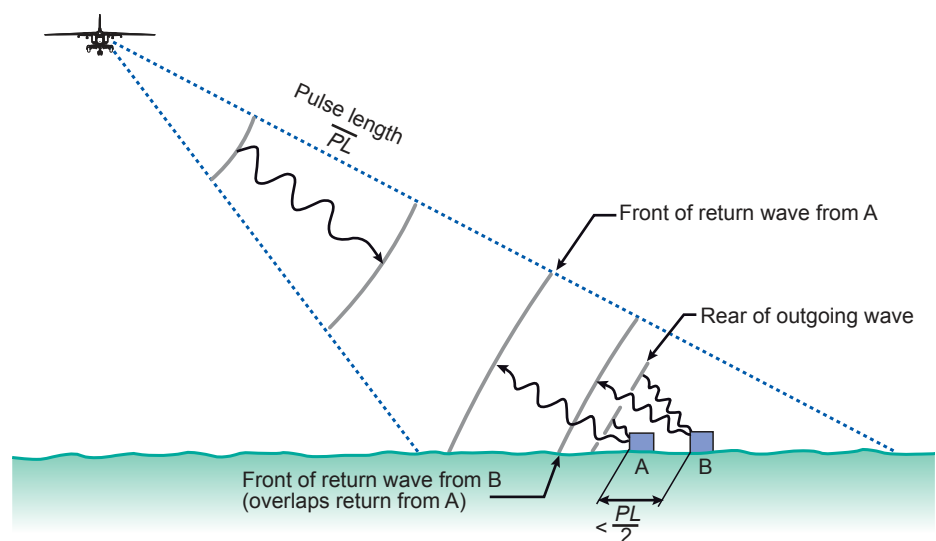
Figure 4.18
Illustration of the slant range resolution.

Scale distortions

Radar measures ranges to objects in slant range rather than true horizontal distances along the ground. Therefore the image has different scales moving from near to far range (Figure 4.17). This means that objects in near range are compressed as compared to objects in far range. For proper interpretation, the image has to be corrected and transformed into ground range geometry.
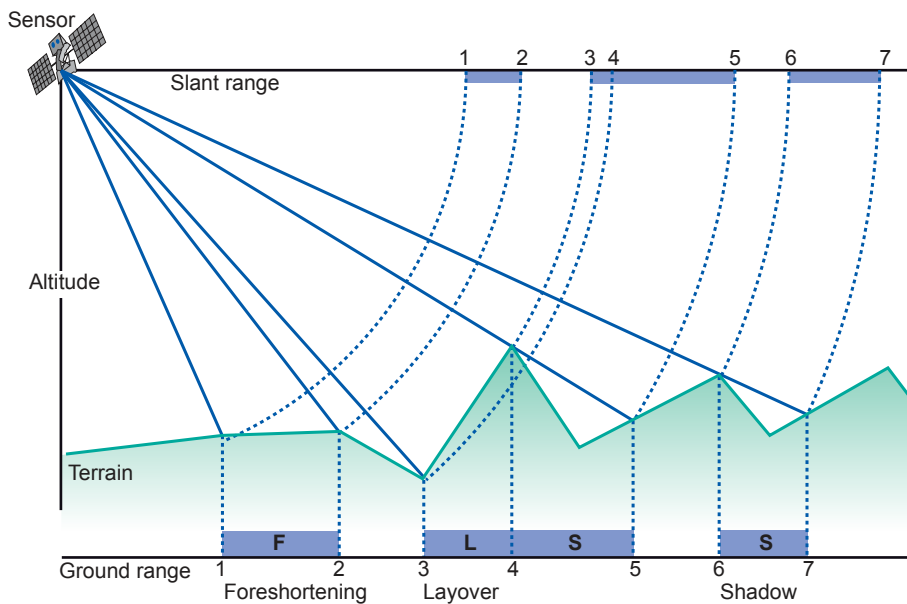
Figure 4.19
Geometric distortions in a radar image caused by varying terrain elevation.

Terrain-induced distortions

Similarly to optical sensors that can operate in an oblique manner (e.g. SPOT), radar images are subject to relief displacements. In the case of radar, these distortions can be severe. There are three effects that are typical for radar: *foreshortening*, *layover* and *shadow* (see Figure 4.19).

Radar measures distance in slant range. The slope area facing the radar is compressed in the image. The amount of shortening depends on the angle that the slope forms in relation to the incidence angle. The distortion is at its maximum if the radar beam is almost perpendicular to the slope. Foreshortened areas in the radar image are very bright.

foreshortening

If the radar beam reaches the top of the slope earlier than the bottom, the slope is imaged upside down, i.e. the slope "lays over". As you can understand from the definition of foreshortening, layover is an extreme case of foreshortening. Layover areas in the image are very bright.

layover

In the case of slopes that are facing away from the sensor, the radar beam cannot illuminate the area at all. Therefore, there is no radiation that can be backscattered to the sensor and so those regions remain dark in the image.

shadow

Radiometric distortions

Geometric distortions also influence the received radiation. Since backscattered radiation is collected in slant range, the received radiation coming from a slope facing the sensor is stored in a reduced area in the image, i.e. it is compressed into fewer pixels

than should be the case if obtained in ground range geometry. This results in high digital numbers because the radiation collected from different objects is combined. Slopes facing the radar appear bright. Unfortunately this effect cannot be corrected for. This is why especially layover and shadow areas in a radar image cannot be used for interpretation. However, they are useful in the sense that they contribute to a three-dimensional appearance of the image and therefore contribute to an understanding of surface structure and terrain relief.

speckle

A typical property of radar images is *speckle*, which appears as grainy "salt and pepper" effects in the image (Figure 4.20). Speckle is caused by the interference of backscattered signals coming from an area that is encapsulated in one pixel. The wave interactions are called *interference*. Interference causes the return signals to be extinguished or amplified, resulting in dark and bright pixels in the image, even when the sensor observes a homogeneous area. Speckle degrades the quality of the image and makes the interpretation of radar images difficult.

interference
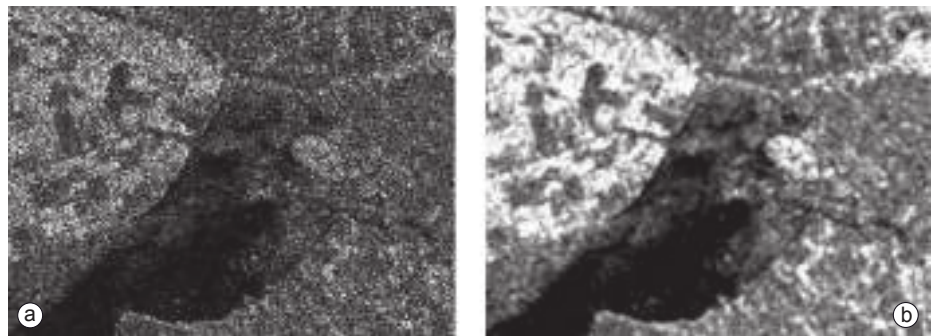
Figure 4.20
An original (a) and speckle filtered (b) radar image.

Speckle reduction

It is possible to reduce speckle by multi-look processing or spatial filtering. If you purchase an ERS SAR scene in "intensity (PRI)-format" you will receive a 3-look or 4-look image. Another way to reduce speckle is to apply spatial filters to the images. Speckle filters are designed to adapt to local image variations in order to smooth values, thus reducing speckle and enhancing lines and edges to maintain the sharpness of an image.

4.4.5   Interpretation of radar images

The brightness of features in a radar image depends on the strength of the backscattered signal. In turn, the amount of radiation that is backscattered depends on a number of factors. An understanding of these factors helps with the proper interpretation of radar images.

Microwave signal and object interactions

For those who are concerned with the visual interpretation of radar images, the degree to which they are able to interpret an image depends upon whether they can identify typical/representative tones related to surface characteristics. The amount of radiation that is received at the radar antenna depends on the illuminating signal (radar system variables such as wavelength, polarization and viewing geometry) and the characteristics of the illuminated object (e.g. roughness, shape, orientation, dielectric constant).

Surface roughness   is the terrain property that most strongly influences the strength of radar backscatter. It is determined by textural features comparable to the size of the radar wavelength (typically between 5 and 40 cm), for example, leaves and twigs of vegetation and sand, gravel and cobble stones. A distinction should be made between surface roughness and terrain relief. Surface roughness occurs at the level of the radar wavelength (centimetres to decimetres). By terrain relief we mean the variation of elevation of the ground surface; relative to the resolution of radar images, only elevation change in the order of metres is relevant. *Snell's law* states that the angle of reflection is equal and opposite to the angle of incidence. A smooth surface reflects the radiation away from the antenna without returning a signal, thereby resulting in a black image. With an increase in surface roughness, the amount of radiation reflected away is reduced and there is an increase in the amount of signal returned to the antenna. This is known as the backscattered component. The greater the amount of radiation returned, the brighter the signal is shown on the image. A radar image is, therefore, a record of the backscatter component and is related to surface roughness.

surface roughness

terrain relief

Complex dielectric constant   Microwave reflectivity is a function of the complex dielectric constant, which is a measure of the electrical properties of surface materials. The *dielectric constant* of a medium consists of a part referred to as permittivity and a part referred to as conductivity [112]. Both properties, permittivity and conductivity, are strongly dependent on the moisture or liquid-water content of a medium. Material with a high dielectric constant has a strongly reflective surface. Therefore the difference in the intensity of the radar return for two surfaces of equal roughness is an indication of the difference in their dielectric properties. In the case of soils, this could be due to differences in soil moisture content.

Surface Orientation   Scattering is also related to the orientation of an object relative to the radar antenna. For example the roof of a building appears bright if it faces the antenna and dark if the incoming signal is reflected away from the antenna. Thus backscatter depends also on the local incidence angle.

Volume scattering   is related to multiple scattering processes within a group of objects, such as the vegetation canopy of a wheat field or a forest. The cover may be all trees, as in a forested area, which may be of different species, with variations in leaf form and size; or grasses and bushes with variations in form, stalk size, leaf and angle, fruiting and a variable soil surface. Some of the radiation will be backscattered from the vegetation surface, but some, depending on the characteristics of radar system used and the object material, will penetrate the object and be backscattered from surfaces within the vegetation. Volume scattering is therefore dependent upon the heterogeneous nature of the object surface and the physical properties of the object, as well as the characteristics of the radar used, such as wavelength and its related effective penetration depth [8].

Point objects   are objects of limited size that give a very strong radar return signal. Usually, the high level of backscatter is caused by *corner reflection*. An example of this is the dihedral corner reflector—a point object situation resulting from two flat surfaces intersecting at 90 ° and situated orthogonally to the incident radar beam. Common forms of dihedral configurations are man-made features such as transmission towers, railway tracks or the smooth side of buildings on a smooth ground surface. Another type of point object is a trihedral corner reflector, which is formed by the intersection of

corner reflection

three mutually perpendicular flat surfaces. Point objects that are corner reflectors are commonly used to identify known fixed points in an area in order to perform precise calibration measurements. Such objects can occur and are best seen in urban areas, where buildings can act as trihedral or dihedral corner reflectors. These objects give rise to intense bright spots on a radar image and are typical for urban areas. Point objects are examples of objects that are sometimes below the resolution of a radar system but, because they dominate the return signal from a cell, nevertheless give a clearly visible point; they may even dominate the surrounding cells.

4.4.6 Applications of radar

There are many useful applications of radar imaging. Radar data provide information complementary to visible and infrared remote-sensing data. In the case of forestry, radar images can be used to obtain information about forest canopy, biomass and different forest types. Radar images can also be used to distinguish between different types of land cover, e.g. urban areas, agricultural fields and water bodies. In urban areas, radar detects buildings (corner reflectors) and metal constructions, thus allowing the extent of urban areas to be delineated, which is a key observable for urban growth studies. In agricultural crop identification, the use of radar images acquired using different polarization (mainly airborne) is quite effective. It is crucial for agricultural applications to acquire data at a certain moment (season) to obtain the necessary parameters. This is possible because radar can operate independently of weather or light conditions. In geology and geomorphology, the fact that radar provides information about surface texture and roughness plays an important role in lineament detection and geological mapping. Since radar backscatter is sensitive to surface roughness, it helps to discriminate between ice and debris, thus making it potentially suitable for glaciers monitoring studies. Radar also allows the measurement of elevation and change in elevation by a technique called Interferometric SAR (INSAR). Radar has also been successfully applied in hydrological modelling and soil moisture estimation— based on the sensitivity of the microwave to the dielectric properties of the observed surface. The interaction of microwaves with ocean surfaces and ice provides useful data for oceanography and ice monitoring. Radar data is also used for oil-slick monitoring and environmental protection.

4.5 Laser scanning

4.5.1 Basic principles

Laser scanning—in functional terms—can be defined as a system that produces digital surface models. The system comprises an assembly of various sensors, recording devices and software. The core component is the laser mechanism. The laser measures distance, which is referred to as "laser ranging". When mounted on an aircraft, a laser rangefinder measures at very short time intervals the distance to the terrain. - Combining a laser rangefinder with sensors that can measure the position and attitude of the aircraft (GPS & IMU) makes it possible to create a model of the terrain surface in terms of a set of (X, Y, Z) coordinates, following the polar measuring principle; see Figure 4.21.

measuring by three sensors
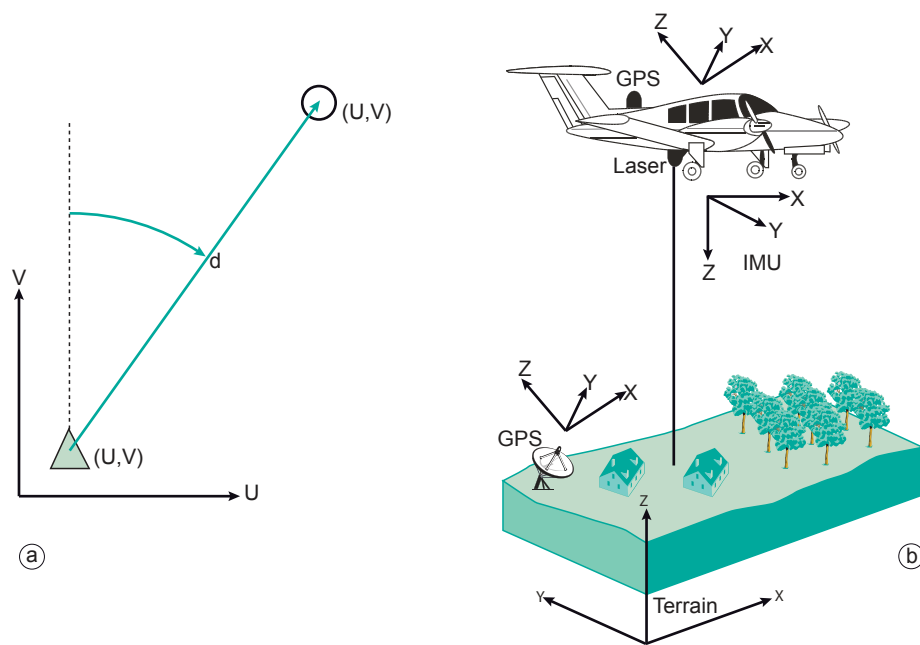
Figure 4.21
Polar measuring principle (a) and its application to ALS (b).

We can define the coordinate system in such a way that Z refers to elevation. The digital surface model (DSM) thus becomes a digital elevation model (DEM), i.e. we model the surface of interest by providing its elevation at many points, each with position coordinates (X, Y). Do the elevation values, which are produced by airborne laser scanning (ALS), refer to elevation of the *bare ground* above a predefined datum? Not necessarily, since the "raw DEM" gives us elevation of the surface the sensor "sees" (Figure 4.22). Post-processing is required to obtain a digital terrain model (DTM) from the DSM.

The key advantages of ALS are its high ranging precision, its ability to yield high resolution DSMs in near real-time, and its complete or nearly complete independence of weather, season or light conditions. Typical applications of ALS are, therefore, forest surveys, surveying of coastal areas and sand deserts, flood-plain mapping, power-line and pipeline mapping, monitoring open-pit mining and 3D city modelling.

applications of ALS
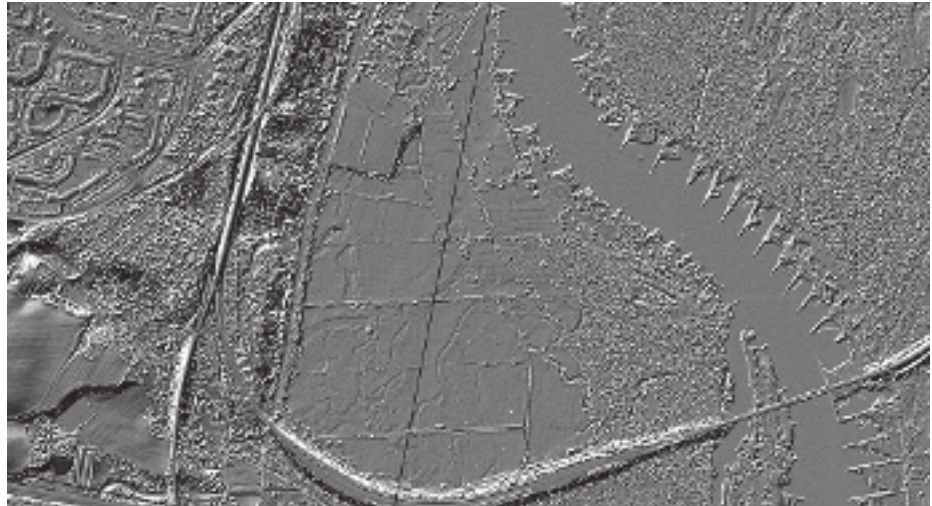
Figure 4.22
DSM of part of Frankfurt (Oder), Germany (1 m point spacing). Courtesy of TopoSys.

4.5.2 ALS components and processes

LASER stands for Light Amplification by Stimulated Emission of Radiation. Although he did not invent it, Einstein can be considered the father of the laser. Roughly 85 years ago he postulated the phenomena of photons and stimulated emission; he won the Nobel prize for related research on the photoelectric effect. In 1960, Theodore Maiman, employed at Hughes Research Laboratories, developed a device to amplify light, thus building the first laser (instrument). A laser emits a beam of monochromatic light or radiation in the NIR range of the spectrum. The radiation is not really of a single wavelength, but it has a very narrow spectral band—smaller than 10 nm. Also specific for lasers is the very high intensity of radiation they emit. Today, lasers are used for many purposes, even human surgery. Lasers can damage cells (by boiling their water content), so they are a potential hazard for the eye. Users of laser rangefinders therefore have to attend safety classes; safety rules must be strictly observed when using lasers for surveying applications.

laser
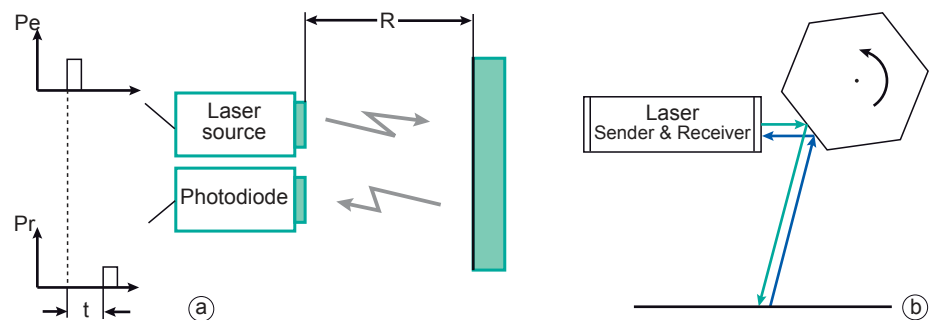
Figure 4.23
Concept of laser ranging and scanning [127].

laser rangefinders

Laser rangefinders and scanners come in various forms. Most airborne laser instruments are "pulse lasers". A pulse is a signal of very short duration that travels as a beam. Airborne laser rangefinders for topographic applications emit NIR radiation. An emitted pulse is reflected by the ground and its return signal is sensed by a photodiode (Figure 4.23). A time counter starts when a pulse is sent out and stops on its return. The elapse time is measured with a resolution of 0.1 ns. As we know the speed of light, c, the elapsed time can easily be converted to a distance:

$$R = \frac{1}{2}\, ct. \qquad\qquad (4.5)$$

Modern laser scanners send out pulses at a very high frequency (up to 300,000 pulses per second). Across-track scanning is in most cases achieved by a moving mirror, which deflects the laser beam. The mirror can be of the oscillating, rotating, or nutating type. The Falcon system uses fiber optics to achieve scanning. Adding a scanning device to a ranging device has made surveying of a large area more efficient: a strip (a swath of points) can be captured on a single leg of , instead of just a line of points as was the case with earlier versions of laser systems (laser profilers).

<div style="float:right">laser scanner</div>

Simple laser rangefinders register one return pulse for every emitted pulse. By contrast, modern laser rangefinders for airborne applications record multiple echoes from the same pulse. Multiple-return laser ranging is specifically relevant for aerial surveys of terrain covered by vegetation because it helps distinguish vegetation echoes from ground echoes. For a pulse that hits a leaf at the top of a tree, part of the pulse may be reflected, while another part of it may travel further, perhaps hitting a branch and, eventually, even the ground; see Figure 4.24. Many of the first return "echoes" will be from the tree canopy, while the last returns are more likely to come from the ground. Each return can be converted to an (X, Y, Z) of the illuminated target point. To figure out whether the point is on the ground or somewhere amongst the vegetation is far from trivial. Multiple return ranging does not give a direct answer but it helps find one. An example of a first return and last return DSM is shown in Figure 4.25. *Full waveform sensors* represent the further development of this approach. Instead of only detecting an echo if its intensity is above a certain threshold (Figure 4.24), full waveform scanners or altimeters (as on ICESat, see below) digitize the entire return signal of each emitted laser pulse. Full waveform laser rangefinders can provide information about surface roughness and more cues on vegetation cover.

<div style="float:right">multiple return ranging</div>

<div style="float:right">full waveform sensors</div>

As well as measuring the range, some laser-based instruments also measure the amplitude of the reflected signal to obtain an image (often referred to as "intensity image"). Imaging by laser scanner is different from imaging by radar instruments. While an image line of a microwave radar image stems from a single pulse, an image line of a laser intensity image stems from many pulses and is formed in the same way as for an across-track multispectral scanner. The benefit of "imaging lasers" is limited. The images obtained are monochromatic and are of lower quality than panchromatic images. A separate camera or multispectral scanner can produce much richer image content.

<div style="float:right">imaging laser</div>

ALS provides 3D coordinates of terrain points. To calculate accurate coordinates of terrain points we must accurately observe all necessary elements. Measuring the distance from the aircraft to the terrain can be done very precisely by the laser rangefinder (accurate to within centimetres), and we can accurately determine the position and altitude of the aircraft using a POS (Section 4.1.1).

<div style="float:right">GPS and IMU</div>

The most widely used platforms for ALS are airplanes and helicopters. Helicopters are better suited for very high-resolution surveys, because they can easily fly slowly. The minimum flying height is, among other things, dependent on the safe eye–laser distance for the instrument. The major limiting factor of the maximum flying height is energy loss of the laser beam. 1000 m and less are frequently used flying altitudes, although there are systems for which heights of 8000 m are feasible.

<div style="float:right">ALS platforms</div>

Unlike aerial surveys for generating stereo coverage of photographs—for which each terrain point should be recorded at least twice—in ALS a terrain point is, in principle, only "collected" once, even if the strips flown overlap. This is an advantage when surveying urban areas and forests, but it has disadvantages for error detection.

After the flight, the recordings from the laser instrument and the POS are co-registered

co-registering the data

to the same time and then converted to (X, Y, Z) coordinates for each point that was hit by the laser beam. The resulting data set may still contain systematic errors and is often referred to as "raw data".

Further data processing has then to solve the problem of extracting information from the un-interpreted set of (X, Y, Z) coordinates. Typical tasks are "extracting buildings", modelling trees (e.g. to compute timber volumes) and, in particular, filtering the DSM to obtain a DTM. Replacing the elevation value at non-ground points by an estimate of the elevation of the ground surface is also referred to as vegetation removal, or "devegging" for short, a term left over from the early days when ALS was primarily used for forested areas (Figure 4.26).

extracting information

Proper system calibration, accurate flight planning and execution (including the GPS logistics), and adequate software are critical factors in ensuring one gets the right data at the right time.

4.5.3 System characteristics

ALS produces a DSM directly comparable with what is obtained by image matching of aerial photographs/images. *Image matching* is the core process of automatically generating a DSM from stereo images. Alternatively, we can also use microwave radar to generate DSMs and—eventually—DTMs. The question is then, why go for ALS? There are in fact several good reasons for using ALS for terrain modelling:

- A laser rangefinder measures distance by recoding the elapse time between emitting a pulse and receiving the reflected pulse from the terrain. Hence, the laser rangefinder is an active sensor and can be used both during daylight hours
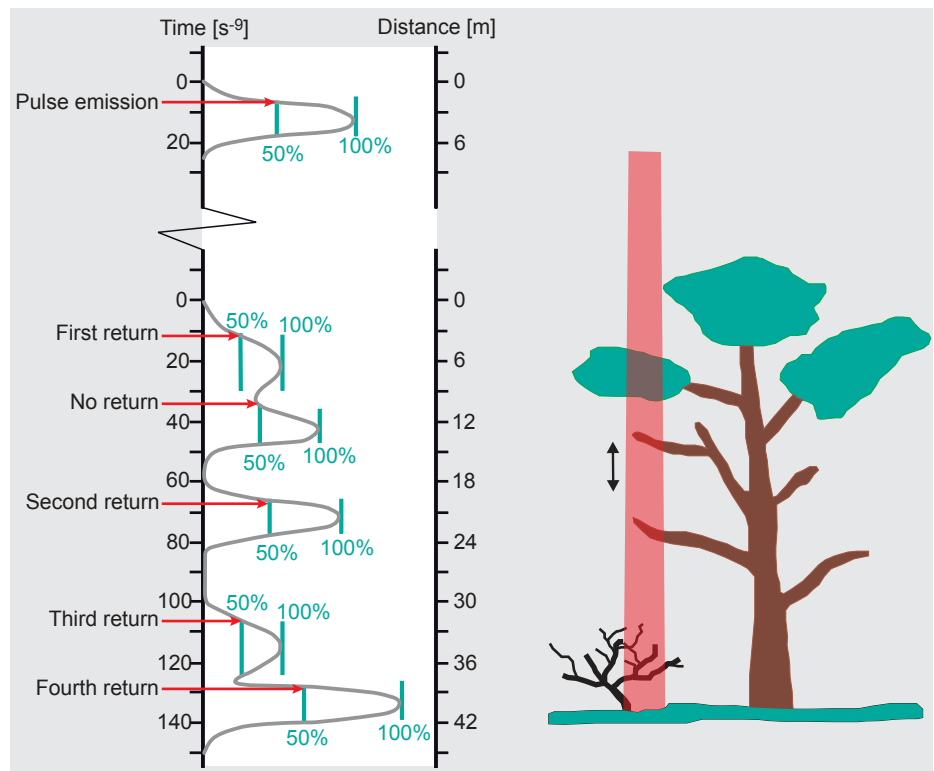
Figure 4.24
Multiple-return laser ranging.
Adapted from Mosaic
Mapping Systems, Inc.
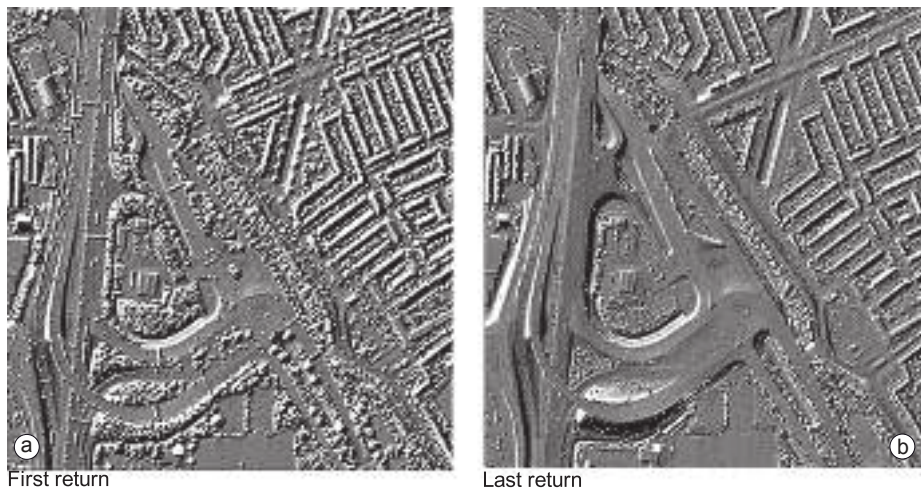
First return

Last return

and at night. The possibility of flying at night comes in handy when, for instance, surveying a busy airport.

- Unlike indirect distance measuring done using stereo images, laser ranging does not depend on surface/terrain texture.

- Laser ranging is less weather-dependent than passive optical sensors. A laser cannot penetrate clouds as microwave radar can, but it can be used at low altitudes, thus very often below the cloud ceiling.

- The laser beam is very narrow, with a beam divergence that can be less than 0.25 mrad; the area illuminated on the ground can, therefore, have a diameter smaller than 20 cm (depending on the laser type and flying height). The simplifying assumption of "measuring points" is thus closely approximated. ALS can "see" objects that are much smaller than the footprint of the laser beam, making it suitable for mapping power lines.

- A laser beam cannot penetrate leaves, but it can pass through the tree canopy, unless that is very dense.

- A laser rangefinder can measure distances very precisely and very frequently; therefore a DSM with a high density of points can be obtained with accurate elevation values. The attainable elevation (vertical coordinate) accuracy with ALS can be in the order of 3 cm for well-defined target surfaces.

- The multiple-return recording facility offers "feature extraction", especially for forest applications and urban mapping (building extraction), both attractive topics for researchers.

- The entire data collection process is digital, which allows it to be automated to a high degree, thus facilitating fast processing.

- Other than a calibration site, which can usually be set up near the airfield, ALS does not need any ground control.

There are two additional major advantages of laser ranging compared to microwave radar: high frequency X pulses can be generated at short intervals and highly directional beams can be emitted. The latter is possible because of the short wavelength of
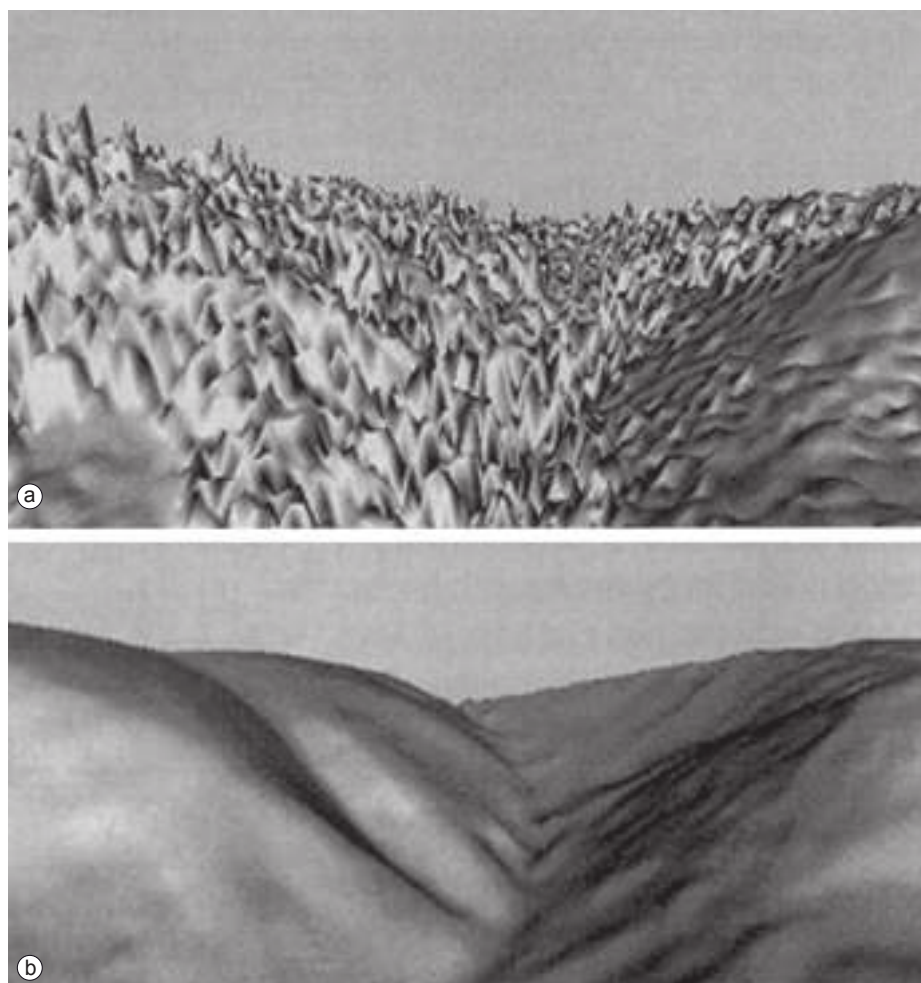
Figure 4.26
Devegging laser data: filtering a DSM (a) to create a DTM (b). From [61].

lasers (10,000 to 1,000,000 times shorter than microwaves). The consequence is much higher ranging accuracy.

Note that the term *radar* is often used as a short form for microwave radar. In the literature, however, you may also come across the term "laser radar", which is synonymous for laser ranging. A more frequently used synonym for laser ranging is LIDAR, although there are also LIDAR instruments that do not measure the distance to but, rather, the velocity of a target ('Doppler LIDARs').

Glacier monitoring   Laser scanning provides information on surface elevation, thus making it a potentially useful tool for monitoring glaciers. However, modern laser scanners usually provide information at very fine spatial resolutions, which are not required in glacier studies. Furthermore, to differentiate glacier ice from debris, we require additional information from other sources.

Urban growth   The sensitivity of laser scanning to the geometric properties of surfaces makes it a suitable tool for detecting objects of urban infrastructure. Laser scan-

ning is, therefore, a potentially useful tool for urbanization studies, especially when very detailed spatial information is required, such as for detecting informal settlements and city construction works. Detailed information about terrain is also relevant for the modelling and monitoring of city growth. Nevertheless, this technique is not currently being used in urban growth studies.

4.6   Aerial photography

Introduction

Aerial photographs have been used since the early 20th century to provide geospatial data for a wide range of applications. *Photography* is the process or art of producing images by directing light onto a light-sensitive surface. Taking and using photographs is the oldest, yet most commonly applied, remote sensing technique. *Photogrammetry* is the science and technique of making measurements on photos and converting these to quantities that are meaningful in the terrain. Some of ITC's early activities included photography and photogrammetry, the latter being, at that time, the most innovative and promising technique available for the topographic mapping of large areas. Aerial film cameras are typically mounted on aircraft, although a Russian satellite is known to have carried a photographic camera and NASA Space Shuttle missions have systematically photographed all aspects of their flights.

Aerial photographs and their digital variant, obtained by digital frame cameras, are today the prime data source for medium- to large-scale topographic mapping and for many cadastral surveys and civil engineering projects, as well as urban planning. Aerial photographs are also a useful source of information for foresters, ecologists, soil scientists, geologists and many others. Photographic film is a very mature medium and aerial survey cameras using film have reached vast operational maturity over the course of many years, so new, significant developments cannot be expected. Owners of aerial film cameras will continue to use them as long as Agfa and Kodak continue to produce film at affordable prices.

Two broad categories of aerial photographs can be distinguished: *vertical* and *oblique* photographs (Figure 4.27). For most mapping applications, vertical aerial photographs are required. A vertical aerial photograph is produced with a camera mounted into the floor of a survey aircraft. The resulting image is similar to a map and has a scale that is roughly constant throughout the image area. Vertical aerial photographs for mapping are usually taken such that they overlap in the flight direction by at least 60%. Two successive photos can form a stereo pair, thus enabling 3D measurement.

vertical photo
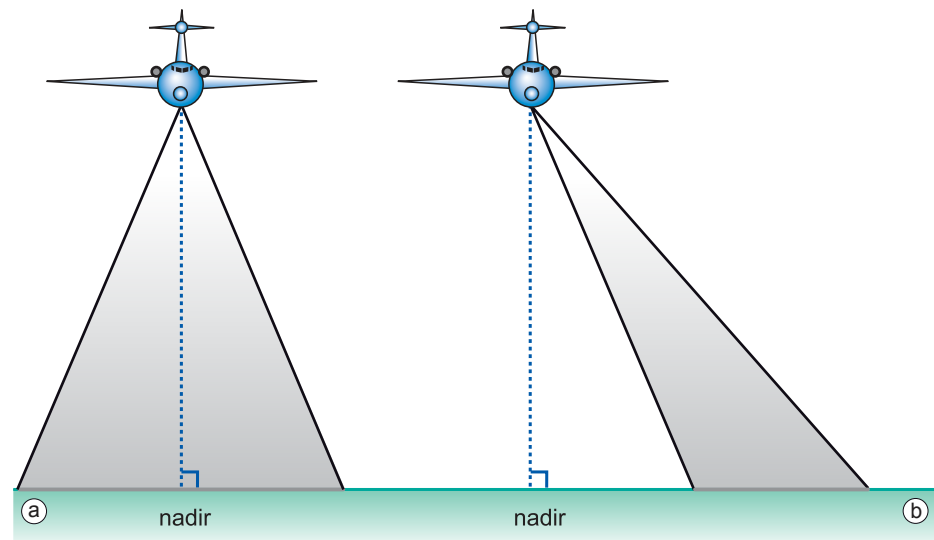
Figure 4.27
Vertical (a) and oblique (b) aerial photography.

Oblique photographs are obtained if the axis of the camera is not vertical. They can

also be made using a hand-held camera and shooting through an open window of an aircraft. The scale of an oblique photo varies from the foreground to the background, which complicates the measurement of positions from the image. For this reason, oblique photographs are rarely used for mapping. Nevertheless, oblique images can be useful for obtaining side views of objects such as buildings.

oblique photo

This section discusses the aerial photo camera, films and methods used for vertical aerial photography. Subsection 4.6.1 describes the aerial camera and its main components. In broad terms, photography is based on the exposure of a photographic film to light, the processing of the film, and the printing of photographs from the processed film. Subsection 4.6.2 discusses the basic geometric—i.e. spatial—characteristics of aerial photographs. Finally, the basics concepts of aerial photography missions are introduced in Subsection 4.6.3.

4.6.1 Aerial survey cameras

A camera used for vertical aerial photography for mapping purposes is called an *aerial survey camera*. Only two manufacturers of aerial survey cameras, namely Leica and Z/I Imaging, have continued to assemble aerial film cameras; their cameras are the RC-30 and the RMK-TOP, respectively. Aerial survey cameras contain a number of components that are also common to any typical hand-held camera, as well as a number of specialized components that are necessary for its specific role. The large size of aerial cameras results from the need to acquire images of large areas with a high spatial resolution. This is achieved by using very large-sized film. Modern aerial survey cameras produce negatives measuring 23 cm × 23 cm (9 inch × 9 inch); up to 600 photographs may be recorded on a single roll of film. To achieve the same degree of quality as an aerial film camera, a digital camera has to produce shots comprising about 200 million pixels.

4.6.2 Spatial characteristics

Two important properties of an aerial photograph are scale and spatial resolution. These properties are determined by sensor (lens cone and film) and platform (flying height) characteristics. Lens cones are available in different focal lengths.

Scale

The relationship between the photo scale factor, s, flying height, H, and focal length, f, is given by

$$s = \frac{H}{f}. \tag{4.6}$$

Obviously, the same scale can be achieved with different combinations of focal length and flying height. If a lens of smaller focal length is used, while the flying height remains constant, then (see also Figure 4.29):

- The *photo scale factor* will increase and the size of individual details in the image will become smaller. In the example shown in Figure 4.29, using a 150 mm and 300 mm lens at $H =$ 2000 m results in scale factors of 13,333 and 6,666, respectively.

- The ground coverage increases. A 23 cm $\times$ 23 cm negative covers an area of 3066 m $\times$ 3066 m if $f = 150$ mm. The width of the coverage reduces to 1533 m if $f = 300$ mm. Subsequent processing takes less time if we can cover a large area with fewer photos.

- The angular field of view increases and the image perspective changes. The FOV for a wide-angle lens is 74°; for a normal angle lens (300 mm) it is 41°. Using shorter focal lengths has the advantage of giving more precise elevation measurements in stereo images (see Section 5.3). Flying a camera with a wide-angle lens at low altitudes has the disadvantage of producing larger obscured areas: if there are tall buildings near the edges of a photographed scene, the areas behind the buildings become hidden because of the central perspective; we call this the *dead ground effect*.
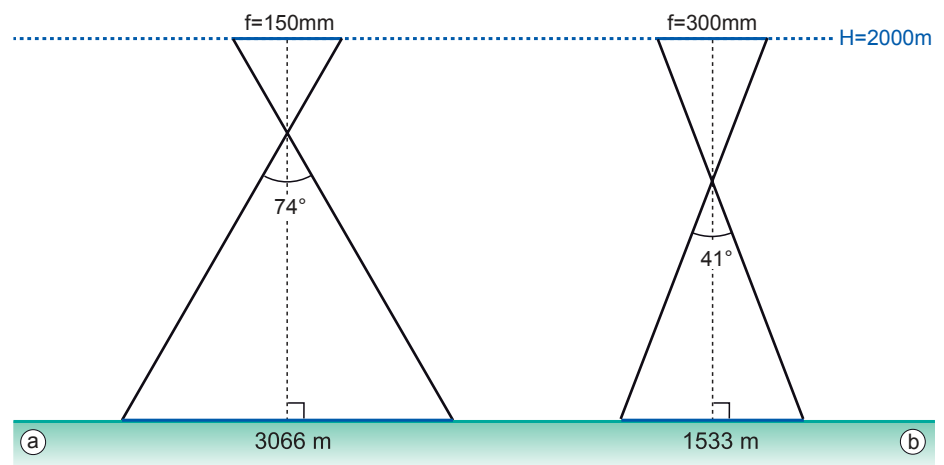
Figure 4.29
The effect of different focal lengths on ground coverage for the same flying height.

Spatial resolution

While scale is a generally understood and applied term, the use of *spatial resolution* in aerial photography is quite difficult. *Spatial resolution* refers to the ability to distinguish small adjacent objects in an image. The spatial resolution of B&W aerial photographs ranges from 40 to 800 line pairs per mm. The better the resolution of a recording system, the more easily the structure of objects on the ground can be viewed in the image. The spatial resolution of an aerial photograph depends on:

- the image scale factor—spatial resolution decreases as the scale factor increases;

- the quality of the optical system—expensive high-quality aerial lenses perform much better than the inexpensive lenses in amateur cameras;

- the grain structure of the photographic film—the larger the grains, the poorer the resolution;

- the contrast of the original objects—the higher the target contrast, the better the resolution,

- atmospheric scattering effects—this leads to loss of contrast and resolution;

- image motion—the relative motion between the camera and the ground causes blurring and loss of resolution.

From this list we can conclude that the actual value of resolution for an aerial photograph depends on quite a number of factors. The most variable factor is the atmospheric conditions, which can change from mission to mission and even during a mission.

4.6.3 Aerial photography missions

Mission planning   When a mapping project requires aerial photographs, some of the first tasks to be done are to select the required photo scale factor, the type of lens to be used, the type of film to be used, and the required percentage of overlap (for stereo pairs). Forward overlap is usually around 60%, while sideways overlap is typically around 20%; Figure 4.30 shows a survey area covered by a number of flight lines. In addition, the date and time of acquisition should be considered with respect to growing season, light conditions and shadow effects.
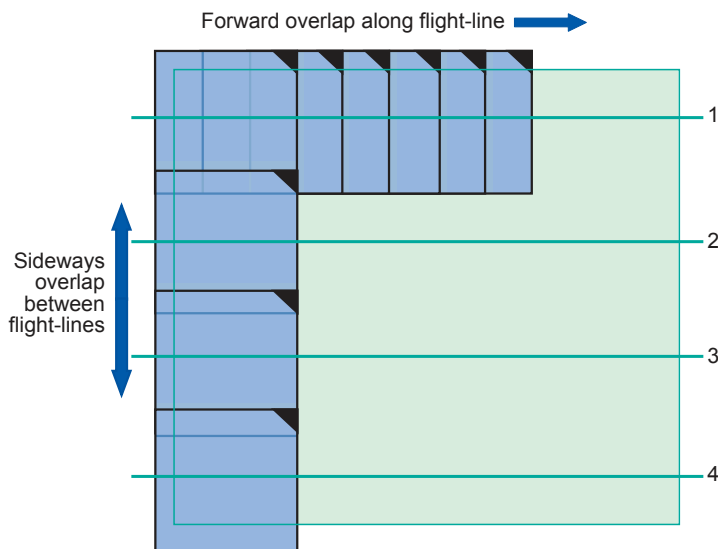
Figure 4.30
Arrangement of photos in a typical *aerial photo block*.

Once the required scale is defined, the following parameters can be determined:

- the required flying height,

- the ground coverage of a single photograph,

- the number of photos required along a flight line,

- the number of flight lines required.

After completion of the necessary calculations, either mission maps are prepared for use by the survey navigator, in the case of a conventional mission execution, or otherwise the data are fed into a mission guidance system.

Mission execution   In current professional practice, we use a computer program to determine, after entering a number of relevant mission parameters and the area of interest, the (3D) coordinates of all positions from which photographs are to be taken. These are stored in a job database. On board, the camera operator/pilot can obtain all relevant information from that database, such as project area, camera and type of film to be used, the number of images required, and constraints regarding time of day or Sun angle, season, and atmospheric conditions.

With the camera positions loaded into a mission guidance system, the pilot is then guided—with the support of GPS—along the mission's flight lines such that deviation from the ideal line (horizontal and vertical) and time to the next exposure station is shown on a display (together with other relevant parameters). If the aircraft passes "close enough" to a predetermined exposure station, the camera fires automatically at the nearest position. This makes it possible to have the data of several projects on board, so that the pilot can choose a project (or part of a project) according to prevailing local weather conditions. If necessary, one can also abandon a project and resume it later.

In the absence of GPS guidance, the aircraft's navigator has to observe the terrain using the traditional viewing system of the aerial camera, check actual flight lines against the planned ones, which are shown graphically on topographic maps, give the required corrections (e.g. to the left or to the right) to the pilot, and tune the overlap regulator to the apparent forward speed of the airplane.

Satellite-based positioning systems and IMU provide a means for achieving accurate navigation. They offer precise positioning of the aircraft during a mission, ensuring that the photographs are taken at the correct points. Computer-controlled navigation and camera management is especially important in survey areas where topographic maps do not exist, are old, or are of small scale or poor quality. They are also helpful in areas where the terrain has few features (sand deserts, dense forests, etc.), because in these cases conventional visual navigation is particularly difficult. The major aerial camera manufacturers (as well as some independent suppliers) now offer complete software packages that enable the flight crew to plan, execute and evaluate an entire aerial survey mission.

4.7 Selection of sensors for a process study

4.7.1 Data selection criteria

For the selection of the appropriate data, it is necessary to fully understand the information requirements of a specific process study. In a nutshell, the questions to be answered concern coverage and resolution in space, time and spectrum. In addition, cost, availability or acquisition constraints, and quality will also be important. The surface characteristics of the object or objects under study determine which parts of the electromagnetic spectrum will be used for observation (spectral coverage), and whether a few broad bands are needed or many narrow bands (spectral resolution).

The level of detail determines the spatial resolution, whereas the size of the area, or the size of the area of the phenomenon, to be studied determines the spatial coverage, which corresponds to the area covered by one image. Of course, one can use several images to cover the area under study, which is often the case, but mosaicking images increases cost and processing time, and it often causes classification and interpretation problems at the seam between two images. Furthermore, the area of the Earth that can be observed by the sensor to be used is an important spatial aspect. For example, the geostationary MSG-SEVIRI covers only the Western Hemisphere, with very large distortions at the poles.

For the temporal aspect, we have to consider the speed of the process and the duration or the length of the period of observation. The speed of the process determines the frequency of observation within a given time (temporal resolution). When choosing the frequency and time of observation and the moments of observation, however, seasonality should be included in the considerations. For example, glaciers shrink in summer and expand in winter. If one wants to study long-term changes in glaciers, (trend versus cyclic changes), images should be recorded at comparable moments in the year, e.g. end-of-winter, at maximum size, end-of-summer, at minimum size, or images at several moments to get a more accurate estimate of seasonal fluctuations, to be able to separate them from long-term trends.

The temporal coverage needed depends on the duration of the process. For the past, temporal coverage is determined by the image archives of a sensor. Landsat archives date back to 1972, but aerial photographs may be available for many decades back. For the future, i.e. both current and planned satellite missions, continuity in the type of sensor are important. Landsat, NOAA, and Meteosat are examples of series of satellites that are each equipped with similar sensors, which guarantees the continuity of data. Security of continuity of data supply is a major issue for many institutes when deciding on which primary data sources to chose. The JERS-1, with its SAR sensor, has long been a typical example of a "one-off" research mission; JERS-1 operated between 1992 and 1998. It was finally followed up in 2006 with the PALSAR sensor on board ALOS.

The selection of data is further influenced by a number of acquisition constraints. Acquisition of optical data is hampered by clouds, so it is not always possible to acquire an image on a planned date, even if the satellite is in the appropriate orbit. Furthermore, not all sensors can record images continuously, because of power limitations, which means that the number of images recorded per orbit is limited. For stereo air photos, occurrence of optimal Sun elevation angles, resulting in enough shadow for the interpretation of height (but not so much that larger parts of the image are obscured), limits the number of days suitable for image acquisition.

Two quality aspects are especially important for process studies: radiometric quality, and calibration. Because of the shorter dwell time per pixel (ground resolution cell),

the radiometric quality of scanners (whiskbroom sensors) is usually less than that of comparable line cameras (pushbroom sensors). Over time, sensors change, so continuous calibration is needed to obtain unbiased observations of the process, and so that trends detected can be attributed to the phenomenon being studied rather than resulting from the aging of the sensor. Furthermore, similar sensors on different platforms in a constellation, or their successors on new platforms, need to be calibrated to make their measurements comparable.

Last but not least, cost plays a major role in image selection for process studies. Although the whole chain of images and processing should be included in cost calculations, in practice the focus tends to be on the cost of the images alone.

To illustrate all these aspects, let us have a look at some typical process studies and the type of data they frequently require. Studies of land processes on a regional or continental scale, for example drought or wildfires, typically use meteorological satellites such as the geostationary Meteosat Second Generation—SEVIRI; the polar orbiting NOAA-AVHRR; or the MODIS sensors of TERRA and AQUA satellites. Spatial resolutions range from 250 m to a few kilometres (depending on location), with the frequency of observations varying from twice a day (MODIS) to every 15 min.

Land use change   Studies of land cover change, deforestation and urban expansion use sensors with spatial resolutions between 15 and 60 m, usually with a temporal coverage of more than a decade; observation once or twice a year is usually sufficient. Detailed change studies, focusing on changes within an urban environment or land cover changes in smaller but fragmented areas, use high resolution sensors. Since data from these sensors, with resolutions ranging from less than a metre up to a few metres, are only available for recent years, they are often combined with older aerial photographs.

Monitoring of glaciers   Glaciers are dynamic objects: their spatial extent is continuously changing. Monitoring of glaciers can be performed daily, monthly, seasonally or yearly. For studies related to global climate change, one would most likely be interested in data of a longer temporal scale. For example, one could select images of the same season for several years in a row. When selecting images, one should keep weather conditions in mind: dense cloud cover or heavy snow covering the land surface will make delineation of the glacier impossible.

Monitoring of Urban growth   The process of urban growth is related to change of land cover type, e.g. construction works, which often take quite some time. Given the typical time scales needed in urban growth studies, annual acquisition of images would be the most appropriate frequency. For studies on urban growth, it does not make sense to acquire images daily.

Chapter 5

Pre-processing

Wan Bakx
Ben Gorte
Wim Feringa
Karl Grabmaier
Lucas Janssen
Norman Kerle
Gabriel Parodi
Anupma Prakash
Ernst Schetselaar
Klaus Tempfli
Michael Weir

5.1 Visualization and radiometric operations

This section explains the processing of raw remote sensing data and the basic principles of visualization of data. The production of images on a computer monitor or paper print-out has always been a prime concern of RS. We use images for inspecting raw data and for performing various data rectification and restoration tasks. Once data are corrected, we convert them once more to images and use these for information extraction by visual interpretation or to support digital image analysis. Many RS applications make use of multispectral data; to visualize them we have to rely on the use of colour. Section 5.1.1 explains how we perceive colour, which can help us to understand how to produce optimal images from multispectral data and how to properly interpret them.

We try to build remote sensors SO that they faithfully image a scene, and we are increasingly successful in doing so. Consider as an example a vertical photograph (or nadir view) of high resolution from a space-borne sensor: it closely resembles a map of the scene and, if the scene was a city, urban planners would be able to readily recognize objects of interest. Taking a closer look, we know that RS images will be geometrically distorted as compared to a map. The degree and type of distortion depends on the type of sensor platform used. Geometrical correction of RS images will be treated in Section 5.3.

In Chapter 2 you have learned that remote sensors measure radiances. The results of

those measurements, however, are recorded as digital numbers, which have no direct physical meaning. The degree to which DNs directly correspond to radiances on the ground depends on many factors. Degradation with respect to what we would like to measure is caused, for example, by unfavourable scene illumination, atmospheric scattering and absorption, and detector-response characteristics. The need to perform radiometric correction in order to compensate for any or all types of degradation depends on the intended use of the data. Urban planners or topographic mappers do not need radiances of objects to be able to recognize them in images. Nevertheless, these images are likely to benefit from "haze correction" and contrast enhancement, to facilitate interpretation. Subsection 5.1.3 therefore briefly treats radiometric correction, only covering corrective measures that are of interest to a wider range of disciplines. (Image restoration and atmospheric correction are discussed further in Section 5.2.) A more detailed description of visualization for map production and spatial analysis is given in Chapter 10.

Elementary image processing techniques to improve the visual quality of an image—so that interpretation becomes easier—are introduced in Section 5.1.4. Image enhancement is not only useful for Earth observation: you may even find it handy for "touching up" your own digital photos.

5.1.1 Visualization

Perception of colour

The perception of colour takes place in the human eye and associated part of the brain. Colour perception concerns our ability to identify and distinguish colours, which in turn enables us to identify and distinguish entities in the real world. It is not completely known how human vision works, or what exactly happens in the eyes and brain before someone decides that an object is, for example, dark blue. Some theoretical models, supported by experimental results, are generally accepted, however. Colour perception theory is applied whenever colours are reproduced, for example in colour photography, TV broadcasting, printing and computer animation.

Tri-stimuli model

We experience light of different wavelengths as different colours. The retinas in our eyes have *cones* (light-sensitive receptors) that send signals to the brain when they are hit by photons that correspond to different wavelengths in the visible range of the electromagnetic spectrum. There are three different kinds of cones, responding predominantly to blue, green and red light (Figure 5.1). The signals sent to our brain by these cones give us sensations of colour. In addition to cones, we have *rods*, which sense brightness. The rods can operate with less light than the cones and do not contribute to colour vision. For this reason, objects appear less colourful in conditions of low light.

cones and rods

This knowledge of the three stimuli is important for displaying colour. Colour television screens and computer monitors are composed of a large number of small dots arranged in a regular pattern of groups of three: a red, a green and a blue dot. At a normal viewing distance from A TV screen, for example, we cannot distinguish the individual dots. We can individually trigger these dots and vary the amount of light emitted from each of them. All colours visible on such a screen are, therefore, created by mixing different amounts of red, green and blue. This mixing takes place in our brain. When we see a mixture of red (say, 700 nm) and green (530 nm) light, we get the same impression as when we see monochromatic yellow light (i.e. with a distinct wavelength of, say, 570 nm). In both cases, the cones are apparently stimulated in the same way. According to the tri-stimuli model, therefore, three different kinds of dots

colour monitors

are necessary and sufficient to recreate the sensation of all the colours of the rainbow.
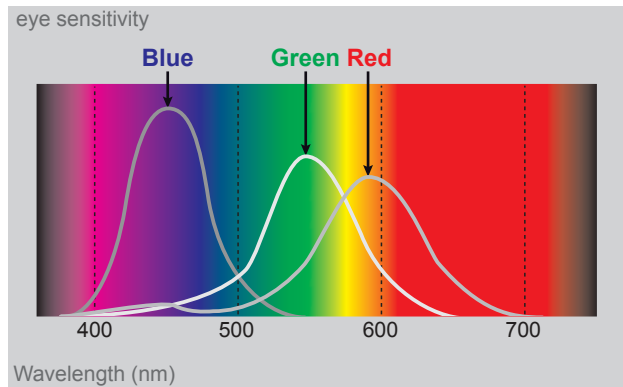
Figure 5.1
Visible range of the
electromagnetic spectrum,
including the sensitivity
curves of cones in the human
eye.

Colour spaces

The tri-stimuli model of colour perception is generally accepted. It states that there are three degrees of freedom in the description of a colour. Various three-dimensional spaces are used to describe and define colours. For our purposes, the following three are sufficient:

1. Red-Green-Blue (RGB) space, which is based on the additive mixing principle of colour;

2. Intensity-Hue-Saturation (IHS) space, which most closely resembles our intuitive perception of colour;

3. Yellow-Magenta-Cyan (YMC) space, which is based on the subtractive principle of colour.

RGB

The RGB definition of colour is directly related to the way in which computer and television screens function. Three channels directly related to the red, green and blue dots are the input to the screen. When we look at the result, our brain combines the stimuli from the red, green and blue dots and enables us to perceive all possible colours from the visible part of the spectrum. During the combination, the three colours are added. We see yellow when green dots are illuminated in addition to red ones. This principle is called the *additive colour scheme*. Figure 5.2 illustrates the additive colours caused by activating red, green and blue dots on a monitor. When only red and green light is emitted, the result is yellow. In the central area, there are equal amounts of light emitted from all three dots, so we experience *white*.

additive colour scheme

In the additive colour scheme, all visible colours can be expressed as combinations of red, green and blue, and can therefore be plotted in a three-dimensional space with R, G and B each being one of the axes. The space is bounded by minimum and maximum values for red, green and blue, thus defining what is known as the colour cube. Figure 5.3 shows the normalized colour cube; the maximum value for each colour is set to 1.

IHS

In day-to-day speech, we do not express colours using the RGB model. The IHS model more naturally reflects our perception of colour. *Intensity* in the colour space describes
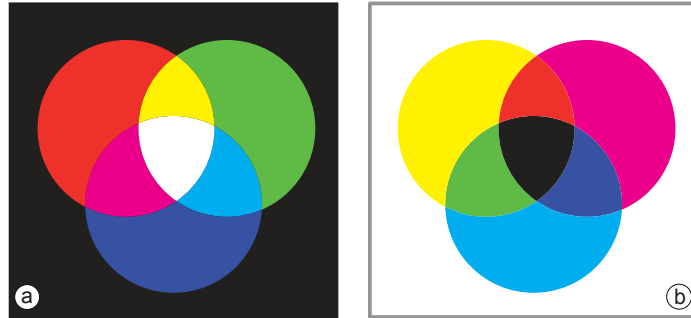
Figure 5.2
Comparison of the
(a) additive and
(b) subtractive colour
schemes.
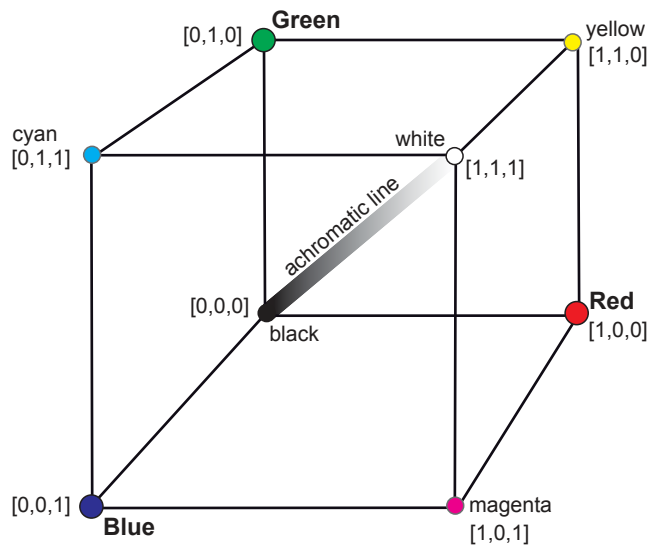
Figure 5.3
The RGB cube. Note the red,
green and blue corner points.

intensity

hue

saturation

whether a colour is dark or light and we use for intensity the value range 0 to 1 (projection on the achromatic diagonal). *Hue* refers to the names that we give to colours: red, green, yellow, orange, purple, etc. We quantify hue by degrees in the range 0 to 360 around the achromatic line. *Saturation* describes a colour in terms of purity and we quantify it as the distance from the achromatic line. "Vivid" and "dull" are examples of common words in the English language that are used to describe colour of high and low saturation, respectively. A neutral grey has zero saturation. As is the case for the RGB system, again three values are sufficient to describe any colour.

Figure 5.4 illustrates the correspondence between the RGB and the IHS models. The IHS colour space cannot easily be transformed to the RGB space because they are completely different. The cube in Figure 5.3 must be converted to a double cone; the inset in Figure 5.4 illustrates this. Although the mathematical model for this description is tricky, the description itself is natural. For example, "light, pale red" is easier to imagine than "a lot of red with considerable amounts of green and blue". The result, however, is the same. Since the IHS model deals with colour perception, which is somewhat subjective, complete agreement of the definitions does not exist. Important for image fusion is the calculation of intensity values and, luckily, this is the simplest of all the calculations. Be aware that the values in the RGB model actually range from 0 to 255, while in the IHS model, intensity ranges from 0 to 1. The formula for intensity is:
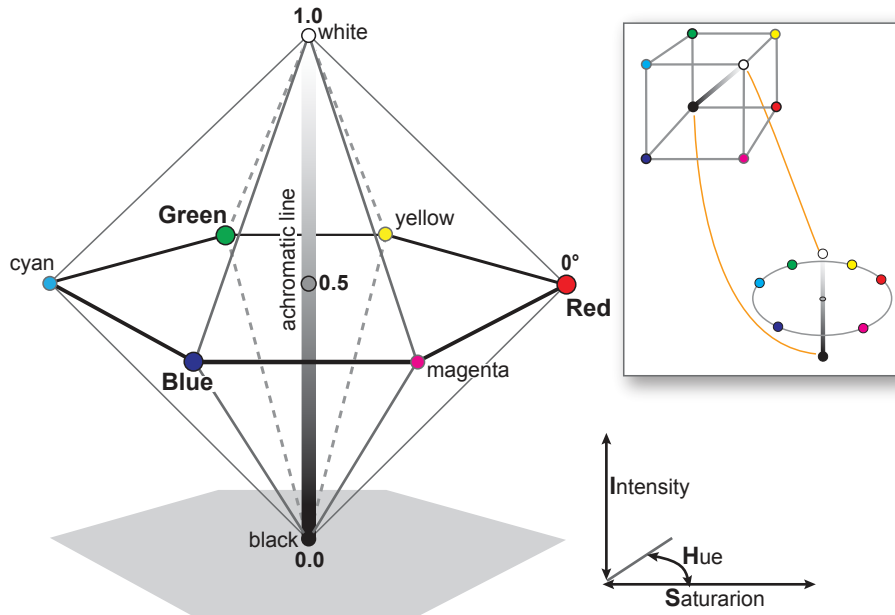
Figure 5.4
Relationship between RGB
and IHS colour spaces.

$$I = \frac{R + G + B}{3 \cdot 255}, \tag{5.1}$$

an orthogonal projection on the achromatic line. For example: $(R, G, B) = (150, 200, 100)$. $I = ((150 + 200 + 100)/(3 \cdot 255)) = 0.59$.

YMC

Whereas RGB is used for computer and TV screens, the YMC colour model is used in colour definition for hardcopy media, such as printed pictures and photographic prints on paper. The principle of YMC colour definition is to consider each component as a coloured filter (Figure 5.2b). The filters are yellow, magenta and cyan. Each filter subtracts one primary colour from the white: the magenta filter subtracts green, so that only red and blue are left; the cyan filter subtracts red, and the yellow filter subtracts blue. Where the magenta filter overlaps the cyan filter, both green and red are subtracted and so we see blue. In the central area, all light is filtered so the result is black. Colour printing, which uses white paper and yellow, magenta and cyan ink, is based on the subtractive colour scheme. When sunlight falls on a colour-printed document, part of it is filtered out by the ink layers and the colour remaining is reflected from the underlying paper.

subtractive colour scheme

5.1.2 Image display

We normally display a digital image using a grey scale. A "digital image" can be raw data such as that obtained with a panchromatic camera, or data obtained by scanning a B&W photograph, or a single band of a multi-band image. For image display, standard computer monitors support 8 bits per pixel. Thus, if we have sensor recordings of 8 bits, each DN will correspond to exactly one grey value. A pixel having the value zero will be shown as black, a pixel having the value 255 as white. Any DN in between becomes, therefore, some shade of grey. One to one correspondence between DN and grey value used to be the standard, so we still often use "grey value" as a synonym for DN. A colour monitor has three input channels, so we have to feed each of them with

grey scale

Figure 5.5
Single-band and three-band image display using the red, green and blue input channels of the monitor.

the same DN to obtain a "grey scale image" (Figure 5.5).

An alternative way of displaying single-band data is to use a colour scale to obtain a *pseudo-colour* image. We can assign colours (ranging from blue via cyan, green and yellow to red) to different portions of the DN range 0–255 (Figure 5.5). The use of pseudo-colour is especially useful for displaying data that are not reflection measurements. With thermal infrared data, for example, the association of cold *versus* warm with blue *versus* red is more intuitive than with dark *versus* bright.

pseudo-colour

When dealing with a multi-band image, any combination of three bands can, in principle, be used as input to the RGB channels of the monitor. The choice should be made based on the intended use of the image. Figure 5.5 indicates how we obtain a *false colour composite*.

colour composites

Sometimes a *true colour composite* is made, where the RGB channels relate to the red, green and blue wavelength bands of a camera or multispectral scanner. An other popular choice is to link RGB to the near-infrared, red and green bands, respectively, to obtain a standard *false colour composite* (Figure 5.6). The most striking characteristic of false colour composites is that vegetation appears as a red-purple colour. In the visible part of the spectrum, plants reflect mostly green light, but their infrared reflection is even higher. Therefore, vegetation displays in a false colour composite as a combination of some blue and a lot of red, resulting in a reddish tint of purple.

true colour

false colour

band 1 (blue)
2 (green)
3 (red)
4 (near-ir)
5 (near-ir)
6 (thermal-ir)
7 (mid-ir)

Natural colour composite (3,2,1)

Pseudo-natural colour composite (3,5,2)

False colour composite (4,3,2)
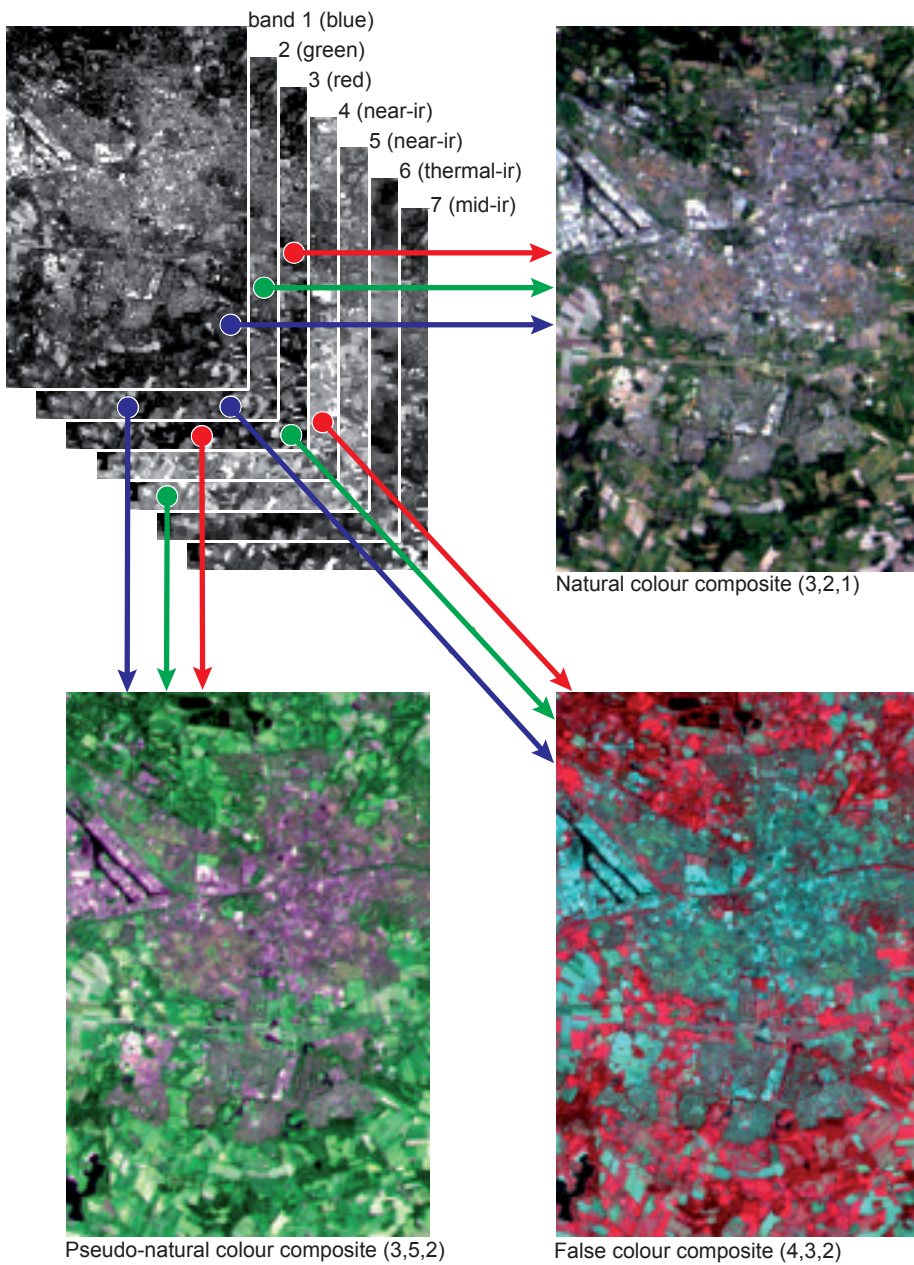
Figure 5.6
Landsat-5 TM false colour
composite of Enschede and
surroundings. Three different
colour composites are
shown: true colour,
pseudo-natural colour and
false colour composites.

Depending on the application, band combinations other than true or false colour may be used. Land use categories can often be distinguished quite well by assigning a combination of Landsat-5 TM bands 5–4–3 or 4–5–3 to RGB.

pseudo-natural colour

Combinations that display NIR as green show vegetation in a green colour and are, therefore, often called *pseudo-natural colour composites* (Figure 5.6). Note that there is no common consensus on the naming of certain composites ("true" may also be referred to as "natural"; "false colour" may also be used for other band combinations than green, red and NIR, etc.). Once you have become familiar with the additive mixing of the red, green and blue primaries, you can intuitively relate the colours—which you perceive on the computer monitor—to the digital numbers of the three input bands, thereby gaining a qualitative insight into the spectral properties of an imaged scene.

anaglyph stereograph

stereoscope

To obtain a 3D visual model from a stereo-image pair on a computer screen, we must combine the images into a stereograph (Figure 5.7) and then use a device that helps us to view the left image with the left eye and the right image with the right eye. There are various technical solutions for this problem, one of which is the *anaglyph* method. The left image is displayed in red, the right one in cyan and the two images are superimposed. For viewing, you need spectacles with a red glass for the left eye and a cyan glass for the right eye. High-end digital photogrammetric systems use polarization instead of colour coding. Polarized spectacles make the images visible to the appropriate eye. The advantage of using polarized images is that we can display a full-colour stereo model and superimpose the results of measurements in any colour. Yet another approach is to use a "split screen" display and a stereoscope in front of the monitor. A stereoscope is a device consisting of a pair of binoculars and two mirrors, which allows two images positioned next to each other to be viewed simultaneously, thus achieving stereoscopic vision. Stereoscopes can also be used to view paper prints of stereo photographs.
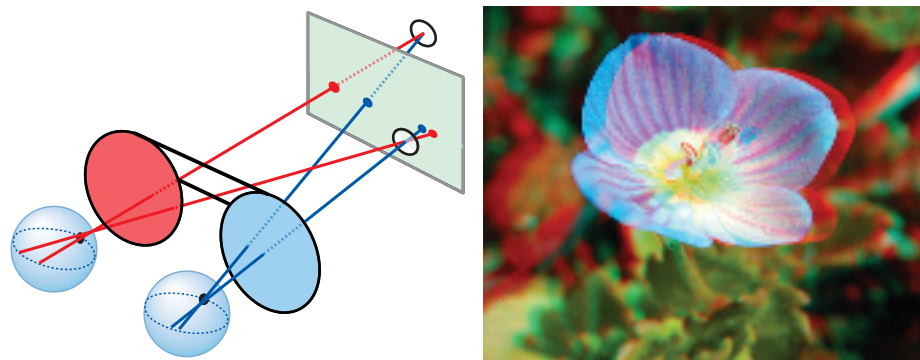
Figure 5.7
The Anaglyph principle and a stereograph.

5.1.3 Radiometric correction

Various techniques can be group under the heading radiometric correction, which aims to correct for various factors that cause degradation of raw RS data. Radiometrically correcting data should make them more suitable for information extraction. Techniques for modifying recorded DNs serve any of the three main purposes outlined below:

image enhancement

- Enhancing images so that they are better suited for visual interpretation. Image enhancement techniques are introduced in a separate section because they can be taken a step further, namely to "low-level image processing" for computer visualization.

- Correcting data for imperfections of the sensor. The detectors of a camera all have a slightly different response. We can determine the differences by radiometric calibration and, accordingly, apply radiometric correction later to the recordings of the camera. Scanners often use several detectors per channel instead of only one. Again, the detectors will each have (slightly) different radiometric responses, with the consequence that the resulting image may be striped. A *destriping* correction will normalize the detectors relatively, if calibration data is absent. A detector may also fail. We may then obtain an image in which, for example, every 10th line is black. A *line drop* correction will cosmetically fix the data. Another detector problem is random noise, which degrades radiometric information content and makes an RS image appear as if salt and pepper was sprinkled over the scene. Correcting all of these disturbances is fairly simple and explained in more detail in Subsection 5.1.8. There can be other degradations caused by the sensor-platform system that are not so easily corrected, such as compensating for image motion blur, which relies on a mathematically complex technique. We got used to referring to theses types of radiometric corrections as *image restoration*. Luckily, image restoration of new sensor data is usually done by the data providers, so you may only have to apply techniques such as destriping and dropped line correction when dealing with old data, e.g. from Landsat MSS. Image restoration should be applied before other corrections and enhancements.

  image restoration

- Correcting data for scene peculiarities and atmospheric disturbances. One scene peculiarity is how the scene is illuminated. Consider an area at an appreciable latitude, such as the Netherlands. The illumination of the area will be quite different in winter than in summer (overall brightness, shadows, etc.), because of differences in Sun elevation. Normalizing images taken in different seasons to make them comparable is briefly outlined below. An atmospheric degradation effect, which is already disturbing when extracting information from one RS image, is atmospheric scattering. Sky radiance at the detector causes haze in the image and reduces contrast. Haze correction is briefly described below. - Converting DNs to radiances on the ground (Section 5.2) becomes relevant if we want to compare RS data with ground measurements, or if we want to compare data acquired at different times by different sensors to detect change.

  scene normalization

  atmospheric correction

5.1.4 Elementary image enhancement

There are two approaches to elementary image processing to enhance an image: histogram operations and filtering. Histogram operations aim at global contrast enhancement, in order to increase the visual distinction between features, while filter operations aim at enhancing local contrast (edge enhancement) and suppressing unwanted image detail. Histogram operations look at DN values without considering where they occur in the image and assign new values from a look-up table based on image statistics. Filtering is a "local operation" in which the new value of a pixel is computed based on the values of the pixels in the local neighbourhood. Figure 5.8 shows the effect of contrast enhancement and edge enhancement for the same input image. Subsection 5.1.5 first explains the notion of histograms.

5.1.5 Histograms

The radiometric properties of a digital image are revealed by its *histogram*, which describes the distribution of the pixel values of the image. By changing the histogram, we change the visual quality of the image. Pixel values (DNs) for 8 bit data range from 0 to 255, so a histogram shows the number of pixels having each value in this

frequency distribution of DNs

175

Figure 5.8
An (a) original, (b) contrast enhanced, and (c) edge enhanced image.

range, i.e. the frequency distribution of the DNs. Histogram data can be represented either in tabular form or graphically. The tabular representation (Table 5.1) usually shows five columns. From left to right these are:

- DN: Digital Numbers, in the range 0–255

- Npix: the number of pixels in the image with a particular DN (frequency)

- Perc: frequency as a percentage of the total number of image pixels

- CumNpix: cumulative number of pixels in the image with values less than or equal to a particular DN

- CumPerc: cumulative frequency as a percentage of the total number of image pixels

Figure 5.9 shows a plot of the columns 3 and 5 of Table 5.1 against column 1. More commonly, the histogram is displayed as a bar graph rather than as a line graph. The graphical representation of column 5 can readily be used to find the '1% value' and the '99% value'. The 1% value is the DN, below which only 1% of all the values are found. Similarly, there are only 1% of all the DNs of the image larger than the 99% value. The 1% and 99% values are often used in histogram operations as cut-off values for display, thus classifying very small and very large DNs as noise outliers rather than signals.

1% and 99% cut-off

A histogram can be "summarized" by descriptive statistics: mean, standard deviation, minimum and maximum, as well as the 1% and 99% values (see Table 5.2). The mean is the average of all the DNs of the image; note that it often does not coincide with the DN that appears most frequently (compare Table 5.1 and Table 5.2). The standard deviation indicates the spread of DNs around the mean.

A narrow histogram (thus, a small standard deviation) represents an image of low contrast, because all the DNs are very similar and are initially mapped to only a few grey values. Figure 5.11a shows the histogram of the image shown in Figure 5.8a. Notice the peak at the upper end (DN larger than 247), while most of DNs are smaller than 110. The peak for the white pixels stems from the sky. All other pixels are dark greyish; this narrow part of the histogram characterizes the poor contrast of the image (a Maya monument in Mexico). Remote sensors commonly use detectors with a

poor contrast

DN	Npix	Perc	CumNpix	CumPerc
0	0	0.00	0	0.00
13	0	0.00	0	0.00
14	1	0.00	1	0.00
15	3	0.00	4	0.01
16	2	0.00	6	0.01
51	55	0.08	627	0.86
52	59	0.08	686	0.94
53	94	0.13	780	1.07
54	138	0.19	918	1.26
102	1392	1.90	25118	34.36
103	1719	2.35	26837	36.71
104	1162	1.59	27999	38.30
105	1332	1.82	29331	40.12
106	1491	2.04	30822	42.16
107	1685	2.31	32507	44.47
108	1399	1.91	33906	46.38
109	1199	1.64	35105	48.02
110	1488	2.04	36593	50.06
111	1460	2.00	38053	52.06
163	720	0.98	71461	97.76
164	597	0.82	72058	98.57
165	416	0.57	72474	99.14
166	274	0.37	72748	99.52
173	3	0.00	73100	100.00
174	0	0.00	73100	100.00
255	0	0.00	73100	100.00

Table 5.1
Example of a histogram in a tabular format.

Mean	StdDev	Min	Max	1% value	99% value
113.79	27.84	14	173	53	165

Table 5.2
Statistics summarizing the histogram of the image represented in Table 5.1.

wide dynamic range, so that they can sense under a wide variety of different illumination and emission conditions. These differences, however, are not always present in one particular scene. In practice, therefore, we often obtain RS images that do not exploit the full range of DNs. A simple technique to achieve better visual quality is, then, to enhance the contrast by "grey scale transformation", which yields a histogram stretched over the entire grey range of pixels on the computer monitor (Figure 5.11c).
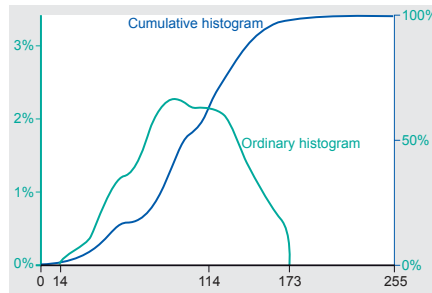
Figure 5.9
Standard histogram and cumulative histogram corresponding to Table 5.1.

5.1.6 Histogram operations

Although contrast enhancement may be done for different purposes, ultimately, in all cases, we will only do so to improve the visual interpretability of an image. In fact, there are two main purposes for applying contrast enhancement. The first is for "temporary enhancement", where we do not want to change the original data but only want to get a better image on the monitor so that we can carry out a specific interpretation task. Image display for geometric correction is an example of this. The second purpose is to generate new data that have a higher visual quality than the original data. This would be the case if an image is to be the output of our RS activities. Orthophoto production and image mapping are examples of such output (see Section 5.3).

Two techniques of contrast enhancement will now be described: *linear contrast stretch* and *histogram equalization* (occasionally also called histogram stretch). Both are "grey scale transformations", which convert input DNs (of raw data) to new brightness values (for a more appealing image) by a user-defined "transfer function".

Linear contrast stretch is a simple grey scale transformation in which the lowest input DN of interest becomes zero and the highest DN of interest becomes 255 (Figure 5.10). The monitor will display zero as black and 255 as white. In practice, we often use the 1% and 99% values as the lowest and highest input DNs. The functional relationship between the input DNs and output pixel values is linear, as shown in Figure 5.11a. The functions shown in the first row (a, d, g) of Figure 5.11 (in the background of the histogram of each input image) are called the *transfer functions*. Many image processing software packages allow users to graphically manipulate the transfer function so that they can obtain an image that appears to their liking. The actual transformation is usually based on a look-up table.

linear contrast stretch

The transfer function to be used can be chosen in a number of ways. We can use linear grey-scale transformation to correct for haze and also for other purposes than contrast stretching, for instance to "invert an image" (convert a negative image to a positive one, or vice versa) or to produce a binary image (the simplest technique of image segmentation). Inverting an image is relevant, for example, after having scanned a negative of a photograph; see the last column of Figure 5.11.

transfer function

Linear stretch is a straight-forward method of contrast enhancement that gives fair results when the input image has a narrow histogram that has a distribution close to
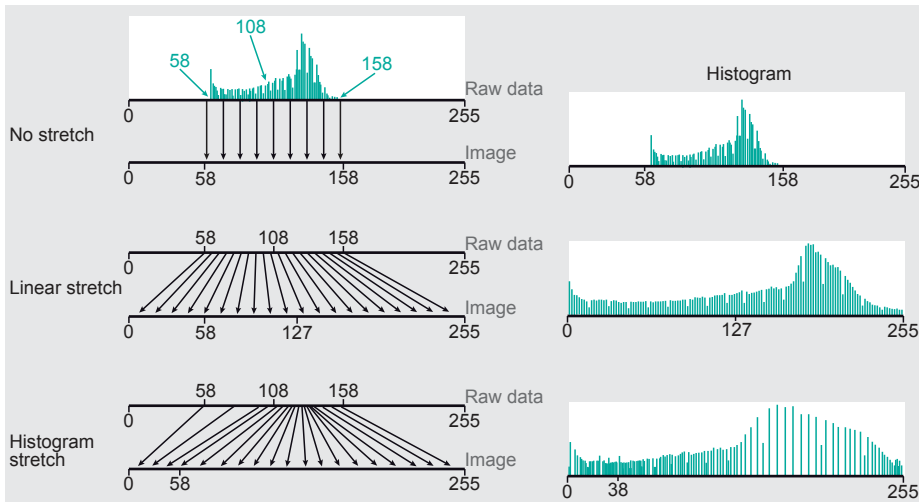
being uniform. The histogram of our example image (Figure 5.8a) is asymmetric, with all DNs in a small range at the lower end, if the irrelevant sky pixels are not considered. Stretching this small range with high frequencies of occurrence (Figure 5.11d) at the expense of compressing the range with only few values (in our example, the range of high brightness) will make more detail (see Figure 5.11e) visible in the dark parts of the original.

histogram equalization

As the name suggests, *histogram equalization* aims at achieving a more uniform distribution in the histogram (see Figure 5.11f). Histogram equalization is a non-linear transformation; several image processing packages offer it as a default function. The idea can be generalized to achieve any desired target histogram.

It is important to note that contrast enhancement (by linear stretch or histogram equalization) merely amplifies small differences between DN values so that we can visually differentiate between features more easily. Contrast enhancement does not, however, increase the information content of the data and it does not consider a pixel neighbourhood. Histogram operations should be based on analysis of the shape and extent of the histogram of the raw data and understanding what is relevant for the intended interpretation. If this is not done, a decrease of information content could easily occur.

5.1.7 Filter operations

A further step in producing optimal images for interpretation is to apply filtering. Filtering is usually carried out for a single band. Filters—algorithms—can be used to enhance images by, for example, reducing noise ("smoothing an image") or sharpening a blurred image. Filter operations are also used to extract features from images, e.g. edges and lines, and to automatically recognize patterns and detect objects. There are two broad categories of filters: linear and non-linear filters.

Linear filters calculate the new value of a pixel as a linear combination of the given values of the pixel and those of neighbouring pixels. A simple example of the use of a linear smoothing filter is when the average of the pixel values in a 3×3 pixel neighbourhood is computed and that average is used as the new value of the central pixel in the neighbourhood (see Figure 5.12). We can conveniently define such a linear filter through a *kernel*. Figure 5.13a shows the kernel of the smoothing filter applied to the example of Figure 5.12. The kernel specifies the size of the neighbourhood that is considered (3×3, or 5×5, or 3×1, etc.) and the coefficients for the linear combination.
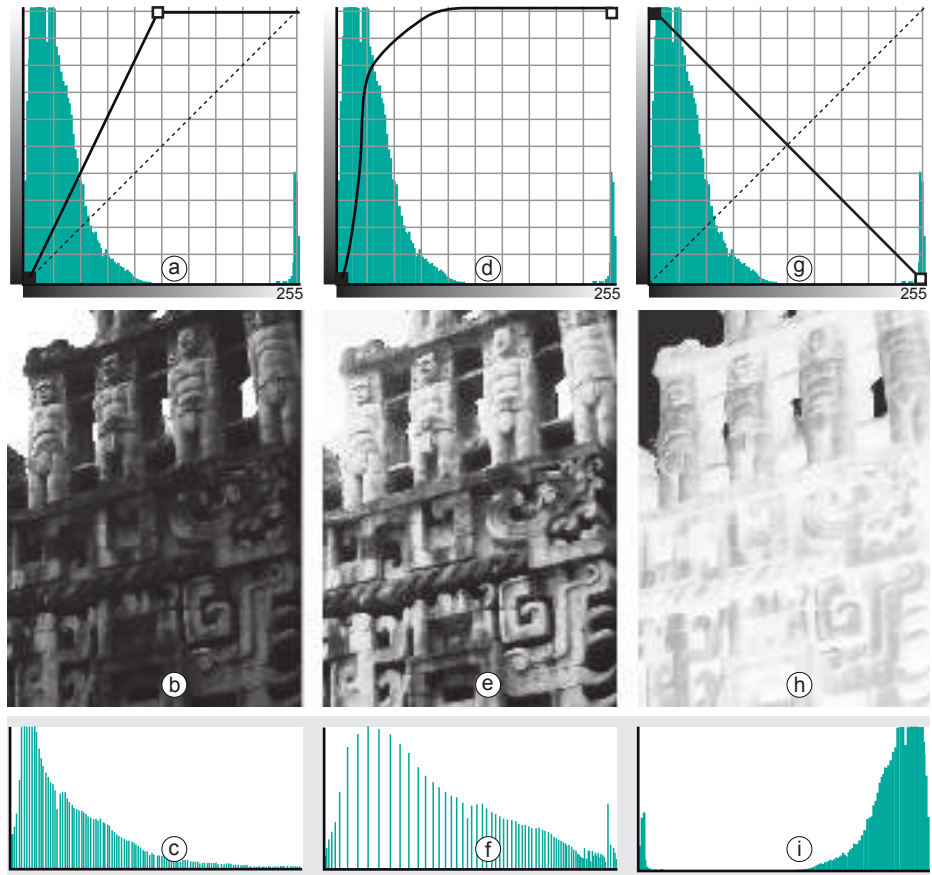
linear filter

Figure 5.11
Effect of linear grey-scale transformations (b, c & h, i) and histogram equalization (e, f).

Image processing software allows us to select a kernel from a list or define our own kernel. The sum of the coefficients for a smoothing filter should be 1, otherwise an unwanted scaling of DN values will result. The filtering software will usually calculate the *gain*

$$gain = \frac{1}{\Sigma k_i} \qquad (5.2)$$

and multiply the kernel values with it. The following two subsections give examples of kernels and their gain. Since there is only one way of using the kernel of a linear filter, the kernel completely defines the filter. The actual filter operation is to "move the kernel over the input image" line by line, thus calculating for each pixel a local combination of pixel values.

moving average

The significance of the gain factor is demonstrated in the following examples. Although in the examples only small neighbourhoods of 3×3 kernels are considered, in practice other kernel dimensions may be used.

Noise reduction

Consider the kernel shown in Figure 5.13a, in which all values equal 1. This means that the values of the nine pixels in the neighbourhood are summed. Subsequently, the result is divided by 9 to ensure that the overall pixel values in the output image
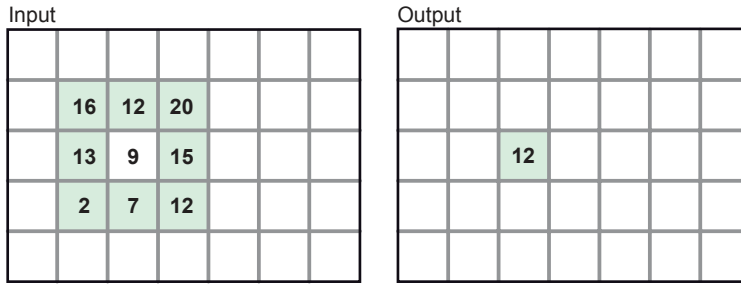
Input

	16	12	20		
	13	9	15		
	2	7	12		

Output

		12			

are in the same range as the input image. In this situation the gain is $1/9 = 0.11$. The effect of applying this *averaging filter* is that an image will become blurred or smoothed. This filter could be applied to radar images to reduce the effect of speckle.

(a)
1	1	1
1	1	1
1	1	1

(b)
1	2	1
2	4	2
1	2	1
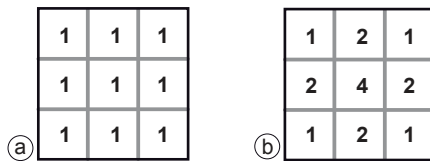
Figure 5.13
Filter kernels for smoothing: (a) equal weights, (b)distance-weighted smoothing.

In the kernel in Figure 5.13a, all pixels contribute equally to the result. It is also possible to define a distance-weighted average instead of an arithmetic mean: the larger the pixel's distance from the centre of the kernel, the smaller the weighting. As a result, less drastic blurring takes place. The resulting kernel, for which the gain is $1/16 = 0.0625$, is given in Figure 5.13b.

Edge detection

Filtering can also be used to detect the edges of objects in images. Such edges correspond to local differences in DN values. This is done using a *gradient* filter, which calculates the difference between neighbour pixels in some direction. Filters presented in Figures 5.14a and b, are called x- and $y-$gradient filters; they perform detection of vertical and horizontal edges, respectively. The filter shown in Figures 5.14c detects edges in all directions. Edge detection filtering produces small values in homogeneous areas of an image, while edges are represented by large positive or negative values. Edge detection filtering can be easily recognized by examining kernel elements: their sum must be zero. This applies to all filters shown in Figure 5.14.

gradient filter

(a)
0	0	0
-1	0	1
0	0	0

(b)
0	-1	0
0	0	0
0	1	0

(c)
-1	-1	-1
-1	8	-1
-1	-1	-1
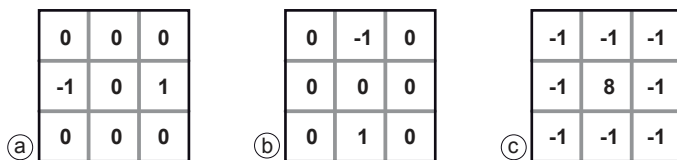
Figure 5.14
Filter kernels for edge detection: (a) x-gradient filter, (b) y-gradient filter, (c) all-directional filter.

Edge enhancement

Filtering can also be used to emphasize local differences in DN values by increasing contrast, for example for linear features such as roads, canals and geological faults. This is done using an *edge enhancing* filter, which calculates the difference between the central pixel and its neighbours. This is implemented using negative values for the non-central kernel elements. An example of an edge enhancement filter is given in Figure 5.15.

-1	-1	-1
-1	16	-1
-1	-1	-1

Figure 5.15
Filter kernel used for edge
enhancement.
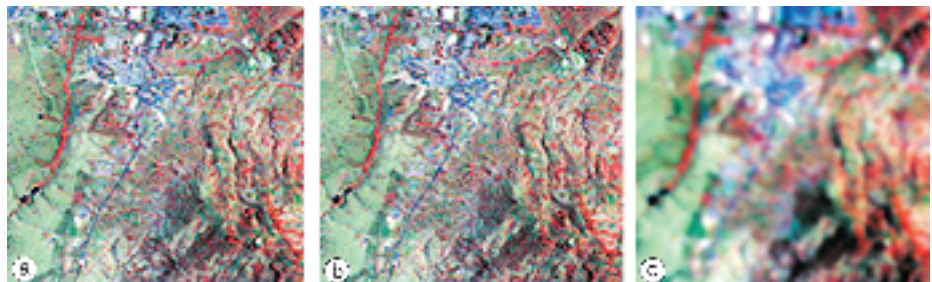
Figure 5.16
Original image (b), edge
enhanced image (a) and
smoothed image (c).

The gain is calculated as: $1/(16 - 8) = 1/8 = 0.125$. The sharpening effect can be made stronger by using smaller values for the centre pixel (with a minimum of 9). An example of the effect of using smoothing and edge enhancement is shown in Figure 5.16.

5.1.8   Correcting data for imperfections of the sensor

The objective of what is called here "cosmetics" is to correct visible errors and noise in the raw data. No atmospheric model of any kind is involved in these correction processes. Instead, corrections are achieved using especially designed filters and image stretching and enhancement procedures. These corrections are mostly executed (if required) at the station receiving the satellite data or at image pre-processing centres, i.e. before reaching the final user. All applications require this form of correction.

Typical problems requiring "cosmetic" corrections are:

- periodic line dropouts;

- line striping;

- random noise or spike.

These effects can be identified visually and automatically; Figure 5.17 illustrates this with Landsat-7 ETM image of Enschede.

Periodic line dropouts

Periodic line dropouts occur due to recording problems when one of the detectors of the sensor in question either gives wrong data or stops functioning. The Landsat-7 ETM, for example, has 16 detectors for each of its channels, except the thermal channel. A loss of one of the detectors would result in every sixteenth scan line being a string of zeros that would plot as a black line on the image (see Figure 5.18).

The first step in the restoration process is to calculate the average DN value per scan line for the entire scene. The average DN value for each scan line is then compared with this scene average. Any scan line deviating from the average by more than a designated threshold value is identified as defective. In regions of very diverse land cover, better results can be achieved by using the histogram for sub-scenes and processing these sub-scenes separately.

The next step is to replace the defective lines. For each pixel in a defective line, an average DN is calculated from the DNs for the corresponding pixel in the preceding and succeeding scan lines by using the principle of spatial autocorrelation. The average DN is then substituted for the defective pixel. The resulting image is a major improvement, although every sixteenth scan line (or every sixth scan line, in the case of Landsat MSS data) consists of artificial data (see Figure 5.19). This restoration program is equally effective for random line dropouts that do not follow a systematic pattern.

Line striping

Line striping is far more common than line dropouts. Line striping often occurs as a result of non-identical detector response. Although the detectors for all satellite sensors are carefully calibrated and matched before the launch of the satellite, with time the response of some detectors may drift to higher or lower levels. As a result, every scan line recorded by that detector is brighter or darker than the other lines (see Figure 5.20). It is important to understand that valid data are present in the defective lines, but these must be corrected to match the overall scene.

Though several procedures can be adopted to correct this effect, the most popular one is histogram matching. Separate histograms corresponding to each detector unit are constructed and matched. Taking one response as standard, the gain (rate of increase of DNs) and offset (relative shift of mean) for all other detector units are suitably adjusted, and new DNs are computed and assigned. This yields a destriped image in which all DN values conform to the reference level and scale.

Random noise or spike noise

Periodic line dropouts and striping are forms of non-random noise that may be recognized and restored by simple means. Random noise, on the other hand, requires a more sophisticated restoration method, such as digital filtering.

Random noise or spike noise may be caused by errors during transmission of data or a temporary disturbance. Here, individual pixels acquire DN values that are much higher or lower than the surrounding pixels (Figure 5.21). In the image, these pixels produce bright and dark spots that interfere with information extraction procedures.

A spike noise can be detected by mutually comparing neighbouring pixel values. If neighbouring pixel values differ by more than a specific threshold margin, it is desig-

96	87	83	94	114	109	107	104	94	83	81	96	90	110	109
89	89	96	94	98	108	111	110	94	80	88	79	73	107	121
104	108	93	92	97	94	86	81	75	83	91	75	75	110	131
114	97	93	86	96	100	92	88	88	92	93	92	105	110	106
131	143	107	101	122	127	88	86	87	90	90	92	124	122	116
122	155	147	115	145	155	101	91	92	95	102	119	131	128	132
112	127	148	151	159	155	134	91	92	91	106	126	121	112	126
137	152	162	164	161	157	150	98	86	96	118	127	104	121	122
138	149	155	159	168	162	154	123	87	97	126	123	108	125	121
131	137	139	141	148	158	159	137	96	92	122	117	114	121	121
108	132	136	134	135	124	115	106	96	88	90	96	113	120	116
88	101	113	100	94	93	90	84	89	79	82	91	88	102	114
78	68	81	88	73	76	76	78	79	72	76	83	86	86	91
70	85	111	118	100	89	86	85	84	81	73	68	73	79	85
87	114	128	127	119	96	87	86	80	74	69	72	76	80	81
106	126	125	127	108	88	92	87	80	80	71	75	77	84	68
106	116	123	119	89	95	99	95	82	76	79	79	79	71	69
99	116	123	119	89	95	99	93	85	74	69	66	70	71	78
88	95	99	96	112	76	104	93	85	74	69	66	70	71	85
91	91	85	87	80	97	115	105	94	87	78	70	73	65	85

Figure 5.17
Original Landsat ETM image of Enschede and and its surroundings (a), and corresponding DNs of the the indicated subset (b).

Figure 5.18
The image with periodic line dropouts (a) and the DNs (b). All erroneous DNs in these examples are shown in bold.

Figure 5.19
The image after correction for line dropouts (a) and the DNs (b).

nated as spike noise and the DN is replaced by an interpolated DN.

96	87	83	94	114	109	107	104	94	83	81	96	90	110	109
89	89	96	94	98	108	111	110	94	80	88	79	73	107	121
104	108	93	92	97	94	86	81	75	83	91	75	75	110	131
150	**127**	**121**	**113**	**129**	**130**	**121**	**116**	**115**	**120**	**127**	**124**	**143**	**148**	**142**
131	143	107	101	122	127	88	86	87	90	90	92	124	122	116
122	155	147	115	145	155	101	91	92	95	102	119	131	128	132
112	127	148	151	159	155	134	91	92	91	106	126	121	112	126
137	152	162	164	161	157	150	98	86	96	118	127	104	121	122
138	149	155	159	168	162	154	123	87	97	126	123	108	125	121
131	137	139	141	148	158	159	137	96	92	122	117	114	121	121
108	132	136	134	135	124	115	106	96	88	90	96	113	120	116
88	101	113	100	94	93	90	84	89	79	82	91	88	102	114
78	68	81	88	73	76	76	78	79	72	76	83	86	86	91
70	85	111	118	100	89	86	85	84	81	73	68	73	79	85
87	114	128	127	119	96	87	86	80	74	69	72	76	80	81
106	126	125	127	108	88	92	87	80	80	71	75	77	84	68
106	116	123	119	89	95	99	95	82	76	79	79	79	71	69
99	116	123	119	89	95	99	93	85	74	69	66	70	71	78
88	95	99	96	112	76	104	93	85	74	69	66	70	71	85
125	**121**	**111**	**119**	**107**	**132**	**160**	**139**	**128**	**121**	**104**	**93**	**99**	**90**	**113**

Figure 5.20
The image with line striping (a) and the DNs (b). Note that the destriped image would look similar to the original image.

96	87	83	94	114	109	107	104	94	83	81	96	90	110	109
89	89	96	94	98	108	111	110	94	80	88	79	73	107	121
104	108	93	92	97	94	86	81	75	**255**	91	75	75	110	131
114	97	93	86	96	100	92	88	88	92	93	92	105	110	106
131	143	107	101	122	127	88	86	87	90	90	92	124	122	116
122	155	147	115	145	155	101	91	92	95	102	119	131	128	132
112	127	148	151	159	155	134	91	92	91	106	126	121	112	126
137	152	162	**0**	161	157	150	98	86	96	118	127	104	121	122
138	149	155	159	168	162	154	123	87	97	126	123	108	125	121
131	137	139	141	148	158	159	137	96	92	122	117	114	121	121
108	132	136	134	135	124	115	106	96	88	90	96	113	120	116
88	101	113	100	94	93	90	84	89	79	82	91	88	102	114
78	68	81	88	73	76	76	78	79	72	76	83	86	86	91
70	85	111	118	100	89	86	85	84	81	73	68	73	79	85
87	114	128	127	119	96	87	86	80	74	69	72	76	80	81
106	126	125	127	108	88	92	87	**255**	80	71	75	77	84	68
106	116	123	119	89	95	99	95	82	76	79	79	79	71	69
99	116	123	119	89	95	99	93	85	74	69	66	70	71	78
88	95	99	96	112	76	104	93	85	74	69	66	70	71	85
91	91	85	87	80	97	115	105	94	87	78	70	73	65	85

Figure 5.21
Image with spike errors (a) and the DNs (b).

5.2   Correction of atmospheric disturbance

Introduction

The radiance values of reflected polychromatic solar radiation and/or the emitted thermal and microwave radiances from a certain target area on the Earth's surface are for researchers the most valuable information obtainable from a remote sensor. In the absence of an atmosphere, the radiance for any wavelength on the ground would be the same as the radiance at the sensor. No atmosphere would make RS easier—but life impossible. So we have to figure out how we can convert remotely detected radiances to radiances at ground level.

In this section we will consider relative and absolute atmospheric correction. Relative atmospheric correction is based on ground reflectance properties, while absolute atmospheric correction is based on atmospheric process information. Before we explain how to correct, Subsection 5.2.1 will review the imaging process and the occurring disturbances.

5.2.1   From satellite to ground radiances: atmospheric correction

The presence of a heterogeneous, dense and layered terrestrial atmosphere composed of water vapour, aerosols and gases disturbs the signal reaching sensors in many ways. Therefore, methods of atmospheric corrections (AC) are needed to "clean" the images from these disturbances, in order to allow the retrieval of pure ground radiances from the target. The physics behind AC techniques in visible and thermal ranges is essentially the same, meaning that the same AC procedures that are applicable to one also apply to the other. However, there are a number of reasons for making a distinction between techniques applied to visible data and thermal data:

- Incident and reflected solar radiation and terrestrial thermal emissions belong to very different parts of the spectrum.

- Solar emission and reflection depends on the position of the Sun and the satellite at the time of image acquisition. Thermal emission is theoretically less dependent on this geometry.

- Solar rays travel twice through the atmosphere before they reach the sensor (Top of the Atmosphere (TOA)–ground–sensor), whereas ground thermal emissions only pass through the atmosphere once (ground–sensor; see Figure 2.6).

- Solar reflection at the Earth's surface depends on material reflectance (ρ). Thermal emission from the Earth depends on the emissivity of the surface materials (ϵ). Since solar reflection and Earth thermal emission occur at different wavelengths, the behaviour of one is not an indication of the other.

- The processes of atmospheric attenuation, i.e. scattering and absorption, are both wavelength dependent and affect the two sectors of the spectrum differently.

- As a result of the previous point, AC techniques are applied at a monochromatic level (individual wavelengths). This means that attenuation of radiation is calculated at every individual wavelength and then integrated across the spectrum of the sensor by mathematical integration.

- Atmospheric components affect different areas of the spectrum in different ways, meaning that some components can be neglected when dealing with data belonging to the thermal or the visible part of the spectrum.

A classification of different AC methods allows us to assess what kind of effort is needed to correct raw data for the particular application at hand. Some RS applications do not require AC procedures at all, except for some "cosmetics", while others call for rigourous and complex procedures. "Intermediate" solutions are sufficient for many applications.

In general, applications for which the actual radiance at ground level is not needed do not require atmospheric correction. Some "cosmetic" and/or image enhancement procedures may suffice: for instance, mapping applications where visual interpretation and image geometry are important, but not the chemical properties of surface material.

Applications that require the quantification of radiation at ground level must include rigourous atmospheric correction procedures. Quantification of evapotranspiration or CO_2 sequestration, or surface temperature and reflectivity mapping, are examples of such applications.

Applications concerned with the evolution of certain parameters or land properties over time, rather than their absolute quantification, are "intermediate" cases. For these, knowledge of the relative trend may suffice. Such procedures apply mainly when the mapping parameters do not really have a meaningful physical value, simply because they were designed primarily for multi-temporal relative comparison. Index evolution and correlation procedures, where radiances are associated with the evolution of a certain parameter (e.g. turbidity) are examples of this category. Be aware that some indexes such as NDVI typically require some absolute atmospheric correction.

The "effort" required is commensurate with the amount of information required to describe the components of the atmosphere at different altitudes (atmospheric profiling) at the moment and position at which the image is taken, and less so with sophistication of the AC procedure itself. State-of-the-art atmospheric models allow the "cleaning" of any cloudless image regardless of sensor type, as long as atmospheric profile data are available. Unfortunately, such detailed atmospheric information can only be obtained through atmospheric sounding procedures, which use a series of instruments to sample the atmosphere at fixed intervals while being transported vertically by a balloon, or sounding sensors on board a satellite. This kind of profiling is carried out daily (at fixed times) at some atmospheric centres, regardless of satellite overpass times. However, the atmosphere is dynamic. Atmospheric processes and composition change rapidly, mainly at low altitudes (water vapour and aerosols), meaning that soundings made somewhere close to the target and near the time of a satellite overpass might not be enough to ensure an adequate atmospheric description. As as rule of thumb regarding AC techniques, first consider the objectives of the project, then identify the appropriate AC procedure, and finally establish the effort, i.e. the required information to execute the chosen correction procedure.

5.2.2 Atmospheric corrections

Haze correction

Equation 2.2 shows that atmospheric scattering adds a "sky radiance". Haze correction aims at removing sky radiance effects from raw data, and doing so can be beneficial to many applications of space-borne RS. In Section 2.4 you have also learned that scattering depends on wavelength: Rayleigh scattering will hardly affect recordings in the red spectral band, while DNs in the blue band may become significantly larger. Reducing haze, therefore, must be done independently for each band of an RS image. How much to subtract from every DN from a particular band? We can find out if the scene is favourable and contains areas that should have zero reflectance (a spectral-band-specific black body). Deep clear water, for example, should yield pixel values of

subtraction per band

zero in the NIR band. If not, we attribute the minimum value found for "water pixels" to sky radiance and subtract this value from all DNs in this band. The alternative is less reliable, i.e. to look at the histogram (see Subsection 5.1.5) of the band and simply take the smallest DN found there as the haze correction constant.

Sun elevation correction

Illumination differences will cause problems if we want to analyse sequences of images of a particular area that were taken on different dates (or images of the same date taken at different time), or if we would like to make mosaics of such images. We can apply a simple Sun elevation correction if the images stem from the same sensor. The trick is to normalize the images as if they were taken with the Sun at its zenith. We can achieve this normalization by dividing every pixel value of an image by the sine of the Sun elevation angle at the time of data acquisition. The Sun elevation angle is usually given in the meta-data file, which is supplied with an image. Obviously this is an approximate correction as it does not take into account the effect of elevation and height differences in the scene, nor atmospheric effects.

normalization by sine

Relative AC methods based on ground reflectance

Relative AC methods avoid the evaluation of atmospheric components of any kind. They rely on the fact that, for one sensor channel, the relation between the radiances at TOA and at ground level follows a linear trend for the variety of Earth features present in the image. This linear relation is in fact an approximation of reality, but for practical purposes it is precise enough when there are other, more important sources of error. The AC methods are:

Two reflectance measurements: The output of this method is an absolute atmospherically corrected image, so it can be used on an individual basis for multi-temporal comparison or parameter evolution and also for flux quantification. "Absolute" means that the image output has physical units and that the calculated ground radiances are compatible with the actual atmospheric constituents. The application of this method requires the use of a portable radiometer able to measure in the same wavelength range as the image band to be corrected. If many bands are to be corrected, then the radiometer should have filters that allow measurement in all these individual bands separately.

Two reference surfaces: The output of this method is an image that matches a reflectance that is compatible with the atmosphere of a similar image taken on a previous date. No absolute values of radiances are obtained in any of the two images, only allowing comparative results. This method works on an individual band/channel basis and is valid for establishing a basis for a uniform comparison to study, for example, the evolution of non-flux related parameters such as indexes, or when certain convenient land properties can be derived directly or indirectly from the normalized radiance values in a band. The method relies on the existence of at least one dark and one bright invariant area. Normally, a sizable area should avoid mixed pixels (mixed land cover). As rule of thumb it should be a minimum of 2 or 3 times larger than the image spatial resolution. Reflective invariant areas are considered to retain their reflective properties over time. Deep reservoir lakes, sandy beaches or deserts, open quarries,large asphalted areas, and large salt deposits are examples of areas that are reflectively invariant. The supposition is that, for these pixels, the reflectance should always be the same since the reflective properties of the materials of which they are composed do not vary with time. If a difference in reflectance occurs for the reflective invariant area in the two date images, it can only be attributed to the different state of the atmosphere on those dates. The atmospheric composition is unknown in the two images, but its influence is measurable by analysing the change in radiance for the reflective invariant areas for the two dates.

Absolute AC methods based on atmospheric processes

These methods require a description of the components in the atmospheric profile. The output of these methods is an image that matches the reflectance of the ground pixels with a maximum estimated error of 10%, provided that atmospheric profiling is adequate enough. This image can be used for flux quantifications, parameter evolution assessments, etc., as mentioned above. The advantage of these methods is that ground reflectance can be evaluated for any atmospheric condition, altitude and relative geometry between the Sun and satellite. The disadvantage is that the atmospheric profiling required for these methods is rarely available. To address this inconvenience, various absolute AC methods have been developed that have different requirements in relation to the atmospheric profiling data—and differences in the accuracy of the results.

Radiative transfer models   Radiative transfer models (RTMs) can be used for computing radiances for a wide variety of atmospheric and surface conditions. They require full descriptions of the atmospheric components at fixed altitudes throughout the atmosphere. RTMs are relatively easy to use when the complexity of the atmospheric input is simplified by using one standard atmosphere as input.

Because of the rapid dynamics of the atmosphere in terms of the temporal and spatial variation of its constituents, researchers have found the need to define some often-observed "common profiles" that correspond to average atmospheric conditions for different parts of the Earth. Compilation of these "fixed atmospheres" has been based on actual radio soundings carried out at different research sites, resulting in what are called "standard atmospheres", e.g. mid-latitude summer, mid-latitude winter, tropical, desert, arctic, US standard, and so on. Researchers use these well-defined standards to characterize typical on-site atmospherics. RTMs have these standards built into the system, allowing the influence of different constituents to be compared under strict simulations. For instance, the influence of water vapour in the thermal, or of aerosols and air molecules in the visible, part of the spectrum can be accurately predicted for different atmospheres, allowing sensitivity analyses for evaluating the importance of these constituents in attenuation processes at different wavelengths.

5.3 Geometric operations

Introduction

If you did not know so before, after reading Chapter 3 you will know that the Earth has a spherical shape and that several clever scientists have devised transformations to map the curved Earth's surface to a plane. Through a map projection (transformation) we can obtain an image of the Earth's surface that has convenient geometric properties. We can, for instance, measure angles on a map and use these for navigation in the real world, or for setting out a designed physical infrastructure. Or if, instead of a conformal projection such as UTM, we use an equivalent projection, we can determine the size of a parcel of land from the map—irrespective of where the parcel is on the map and at which elevation it is on the Earth. A remote sensor "images" the Earth's surface without knowledge of map projections, so we must not expect that remote sensing images have the same geometric properties as a map. Moreover, wherever a remote sensor detects, it merely records DNs. DNs do not come with a label that tell us where exactly on the Earth is the corresponding ground-resolution cell to be found (Figure 5.22).
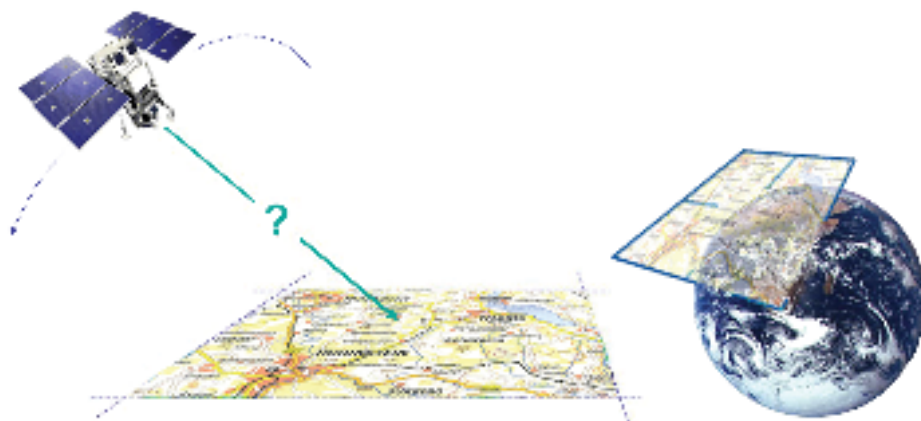
Figure 5.22
The problem of georeferencing an RS image.

Luckily, we know that the DNs are delivered in an orderly fashion, neatly arranged in rows and columns. The position of a point in the image is uniquely defined by the row and column numbers of the pixel that represents the point. Relating a pixel position in the image to the corresponding position on the ground is the purpose of *georeferencing the image*. Once we have figured out what the geometric relationship is between points on the ground and the corresponding point in the RS image, we can transform any pixel position to a position on the ground. "Position on the ground" can be defined either in a 3D terrain coordinate system or through a map projection in a 2D map coordinate system. By georeferencing an image we solve two problems at the same time: (1) we can get map coordinates of features that we identify in the image, and (2) we implicitly correct for geometric distortions of the RS image if we compute correct map coordinates for any pixel. It takes georeferencing to turn RS data into geospatial data. After georeferencing (or, speaking more generally, sensor orientation), we can:

- make measurements within images to obtain 2D and 3D object descriptions. Mapping the world in which we live has been man's concern for thousands of years, with navigation and military activities being the main triggers for topographic mapping. RS—photogrammetry, more specifically—has made that

mapping much more efficient. Mapping is still the prime application of RS, although environmental monitoring is catching up quickly because of the exponential damage we are causing to our environment. We are interested in mapping our environment at a variety of spatial and thematic resolutions and accuracies. For many applications, 2D representations (by points, lines and areas) of objects suffice. As long as certain conditions are met, we can obtain these representations from a single image and simple georeferencing, which directly relates the RS image to the digital map. We need stereo images or multiple images for applications requiring 3D coordinates, or for better 2D coordinates, such as for mapping scenes of large elevation differences or objects with large height differences. Sensor orientation must then use more rigourous approaches than for 2D georeferencing.

- combine an image with other images or vector (digital map) data (see also Chapter 11). Assume you would like to see how land property units relate to land cover. If you had a digital cadastral map, you could readily overlay the parcel boundaries on a RS image, e.g. a scanned photograph, which shows land cover nicely. Combining different RS images and/or map data can be done conveniently if all the data sets do not differ in their geometry, if they are all transformed to the same geometric reference system. From a computational perspective, producing a new image from an RS image such that it fits a specific map projection is often referred to as *geocoding*.

5.3.1 Elementary image distortions

Each sensor–platform combination is likely to have its own type of geometric image distortion. Here we only examine three very common types: (1) the effect of oblique viewing, (2) the effect of Earth rotation, and (3) "relief displacement". Tilting a camera (see Figure 4.27) leads to images of non-uniform scale (Figure 5.23). Objects in the foreground appear larger than those farther away from nadir. Earth rotation affects space-borne scanners and line camera images that have a large spatial coverage. The resulting geometric image distortion (Figure 5.23) can easily be corrected by 2D georeferencing. Relief displacement shows up specifically in large-scale camera images if there is significant terrain relief or if there are high, protruding objects.

oblique view
Earth rotation

Area on the ground

Oblique photo of the area

Area on the ground

Scanner image of flat terrain
on a rotating Earth
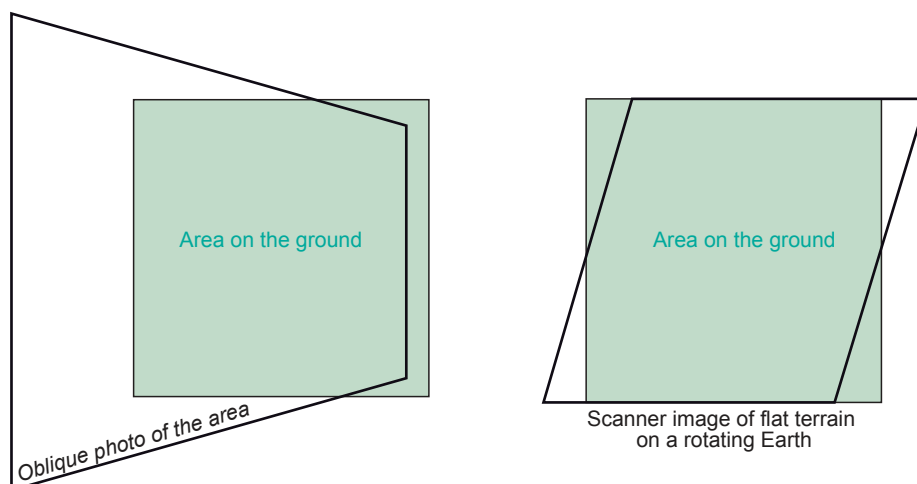
Figure 5.23
Examples of geometric image distortion.

Relief displacement

A characteristic of most sensor systems is the occurrence of distortion of the geometric relationship between the image and a conventional map of the terrain due to elevation differences. This effect is most apparent in aerial photographs but also occurs in images from space-borne line cameras. The effect of relief displacement is illustrated in Figure 5.24 for a line camera. Consider the situation on the left (a), in which a true vertical image is taken of flat terrain. The distances $A - B$ and $a - b$ are proportional to the total width of the scene and its image size, respectively. In the situation on the left, by using the scale factor we can compute $A - B$ from a measurement of $a - b$ in the image. In the situation on the right (b), there is a significant difference in terrain elevation. As you can now observe, the distance between a and b in the image has become larger, although when measured in the terrain system, it is still the same as in the situation on the left. This phenomenon does not occur in the centre of a central projection image but becomes increasingly prominent towards the edges of a camera image. This effect is called *relief displacement*: terrain points whose elevation is above or below the reference elevation are displaced, respectively, away from or towards the nadir point, A, the point on the ground directly beneath the sensor. The magnitude of displacement, δr (in mm), is approximated by:

$$\delta r = \frac{rh}{H}. \tag{5.3}$$

In this equation, r is the radial distance (in mm) from nadir, h (in m) is the terrain elevation above the reference plane, and H (in m) is the flying height above the reference plane (where nadir intersects the terrain). The equation shows that the amount of relief displacement is zero at nadir ($r = 0$) and largest at the edges of a line camera image and the corners of a frame camera image. Relief displacement is inversely proportional to the flying height.

Figure 5.24
Illustration of the effect of terrain topography on the relationship between $A - B$ on the ground and $a - b$ in the image: (a) flat terrain, (b) significant elevation difference.
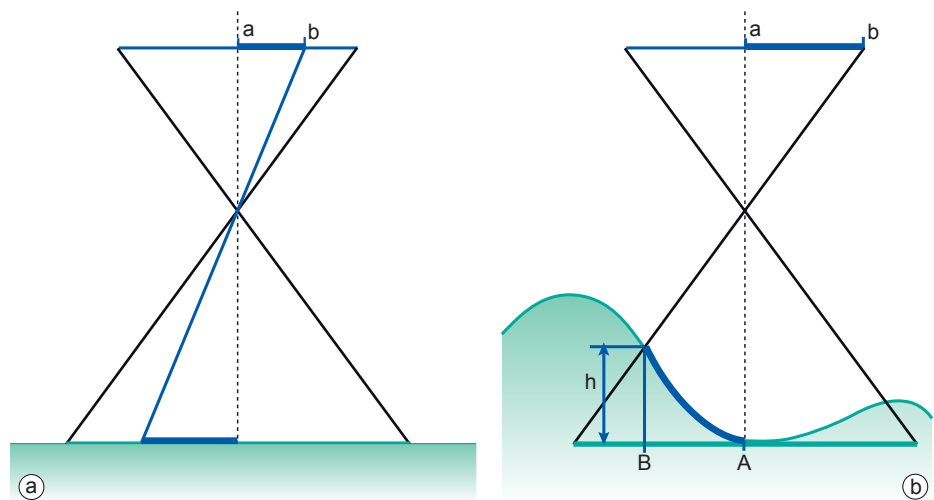
If the scene is just barren land, we cannot see any relief displacement. However, we can see relief displacement if there are protruding objects in the scene (occasionally referred to as height displacement). On large-scale photographs or very high-resolution space-borne images, buildings and trees appear to lean outwards, away from the nadir point (Figure 5.25).

The main effect of relief displacement is that inaccurate or wrong map coordinates will

shifts due to height and elevation

Figure 5.25
Fragment of a large-scale
aerial photograph of the
centre of Enschede. Note the
effect of height displacement
on the higher buildings.

be obtained when, for example, digitizing images without further correction. Whether relief displacement should be considered in the geometric processing of RS data depends on its impact on the accuracy of the geometric information derived from the images. Relief displacement can be corrected for if information about terrain relief is available (in the form of a DTM); see Subsection 5.3.4 for more details. It is also useful to remember that it is relief displacement that allows us to perceive depth when looking at a stereograph and to extract 3D information from such images.

effect of relief displacement

Two-dimensional approaches

This subsection deals with the geometric processing of RS images in situations where relief displacement can be neglected, for example for a scanned aerial photograph of flat terrain. For practical purposes, "flat" may be considered as $h/H < 1/1000$, though this also depends on project accuracy requirements; h stands for relative terrain elevation, H for flying height. For space-borne images of medium spatial resolution, relief displacement is usually less than a few pixels in magnitude and thus less important, as long as near-vertical images are acquired. The objective of 2D georeferencing is to relate the image coordinate system to a specific map coordinate system (Figure 5.26).
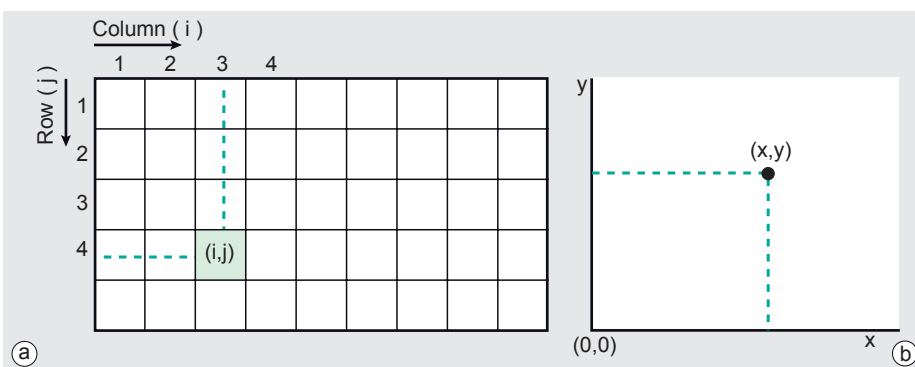
Figure 5.26
Coordinate system of (a) the
image defined by rows and
columns, and (b) the map
with x- and y-axes.

5.3.2 Georeferencing

The simplest way to link image coordinates to map coordinates is to use a transformation formula. A *geometric transformation* is a function that relates the coordinates of two systems. A transformation relating (x, y) to (i, j) is commonly defined by linear equations, such as: $x = 3 + 5i$, and $y = -2 + 2.5j$.

Using the above transformation, for example, the image position $(i = 3, j = 4)$ corresponds to map coordinates $(x = 18, y = 8)$. Once the transformation parameters have been determined, the map coordinates for each pixel can be calculated. This implies that we can superimpose data that are given in the map coordinate system on the image vector, or that we can store features by map coordinates when applying on-screen digitizing. Note that the image in the case of georeferencing remains stored in the original (i, j) raster structure and that its geometry is not altered. As we will see in Subsection 5.3.3, transformations can also be used to change the actual shape of an image and thus make it geometrically equivalent to the map.

type of transformation

The process of georeferencing involves two steps: (1) selection of the appropriate type of transformation, and (2) determination of the transformation parameters. The type of transformation depends mainly on the sensor–platform system used. For aerial photographs (of flat terrain) what is known as "projective transformation" models well the effect of pitch and roll (see Figures 4.27, 5.23, and 5.28). Polynomial transformation, which enables 1st, 2nd to nth order transformations, is a more general type of transformation. In many situations a 1st order transformation is adequate. Such transformation relates map coordinates (x, y) with image coordinates (i, j) as follows:

$$x = a + bi + cj \tag{5.4}$$

$$y = d + ei + fj \tag{5.5}$$

ground control points

Equations 5.4 and 5.5 require that six parameters (a to f) be determined. The transformation parameters can be determined by means of *ground control points* (GCPs). GCPs are points that can be clearly identified in the image and on the target map. The target map could be a topographic map or another image that has been transformed beforehand to the desired map projection system. The operator then needs to identify corresponding points on both images. The image and map scale determine which points are suitable. Typical examples of suitable points are road crossings, crossings of waterways and salient morphological structures. Another possibility is to identify points in the image and to measure the coordinates of these points in the field, for example by GPS, and then transform those to map coordinates. It is important to note that it can be quite difficult to identify good GCPs in an image, especially in lower-resolution space-borne images. Once a sufficient number of GCPs have been specified, software is used to determine the parameters a to f of the Equations 5.4 and 5.5 and quality indications.

number of GCPs

To solve the 1st order polynomial equations, only three GCPs are required; nevertheless, you should use more points than the strict minimum. Using merely the minimum number of points for solving the system of equations would obviously lead to a wrong transformation if you made an error in one of the measurements, whereas including more points for calculating the transformation parameters enables software to also compute the error of the transformation. Table 5.3 gives an example of the input and output of a georeferencing computation in which five GCPs have been used. Each GCP is listed with its image coordinates (i, j) and its map coordinates (x, y).

Software performs a "least-squares adjustment" to determine the transformation pa-

GCP	i	j	x	y	x_c	y_c	d_x	d_y
1	254	68	958	155	958.552	154.935	0.552	-0.065
2	149	22	936	151	934.576	150.401	-1.424	-0.599
3	40	132	916	176	917.732	177.087	1.732	1.087
4	26	269	923	206	921.835	204.966	-1.165	-1.034
5	193	228	954	189	954.146	189.459	0.146	0.459

Table 5.3
A set of five ground control points, which are used to determine a 1st order transformation. x_c and y_c are calculated using the transformation, d_x and d_y are the residual errors.

rameters. The least squares adjustment ensures an overall best fit of the GCPs. We then use the computed parameter values to calculate coordinates (x_c, y_c) for any image point (pixel) of interest:

$$x_c = 902.76 + 0.206i + 0.051j,$$

and

$$y_c = 152.579 - 0.044i + 0.199j.$$

For example, for the pixel corresponding to GCP 1 ($i = 254$, $j = 68$) we can calculate the transformed image coordinates x_c and y_c as 958.552 and 154.935, respectively. These values deviate slightly from the input map coordinates (as measured on the map). Discrepancies between measured and transformed coordinates of GCPs are called residual errors (*residuals* for short). The residuals are listed in the table as d_x and d_y. Their magnitude is an indicator of the quality of the transformation. Residual errors can be used to analyse whether all GCPs have been correctly determined.

residuals

The overall accuracy of a transformation is either stated in the accuracy reportusually provided by software in terms of variances or as *Root Mean Square Error* (RMSE), which calculates a mean value from the residuals (at check points). The RMSE in the x-direction, m_x, is calculated using the following equation:

$$m_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\delta x_i{}^2}. \tag{5.6}$$

For the y-direction, a similar equation can be used to calculate m_y. The overall error, m_p, is calculated by:

$$m_p = \sqrt{m_x{}^2 + m_y{}^2}. \tag{5.7}$$

For the example data set given in Table 5.3, the residuals m_x, m_y and m_p are 1.159, 0.752 and 1.381, respectively. The RMSE is a convenient measure of overall accuracy, but it does not tell us which parts of the image are accurately transformed and which parts are not. Note also that the RMSE is only valid for the area that is bounded by the GCPs. In the selection of GCPs, therefore, points should be well distributed and include locations near the edges of the image.

5.3.3 Geocoding

The previous subsection explained that two-dimensional coordinate systems, e.g. an image system and a map system, can be related using geometric transformations. This georeferencing approach is useful in many situations. However, some situations a *geocoding* approach, in which the row–column structure of the image is also transformed, is required. Geocoding is required when different images need to be combined or when the images are used in a GIS environment that requires all data to be
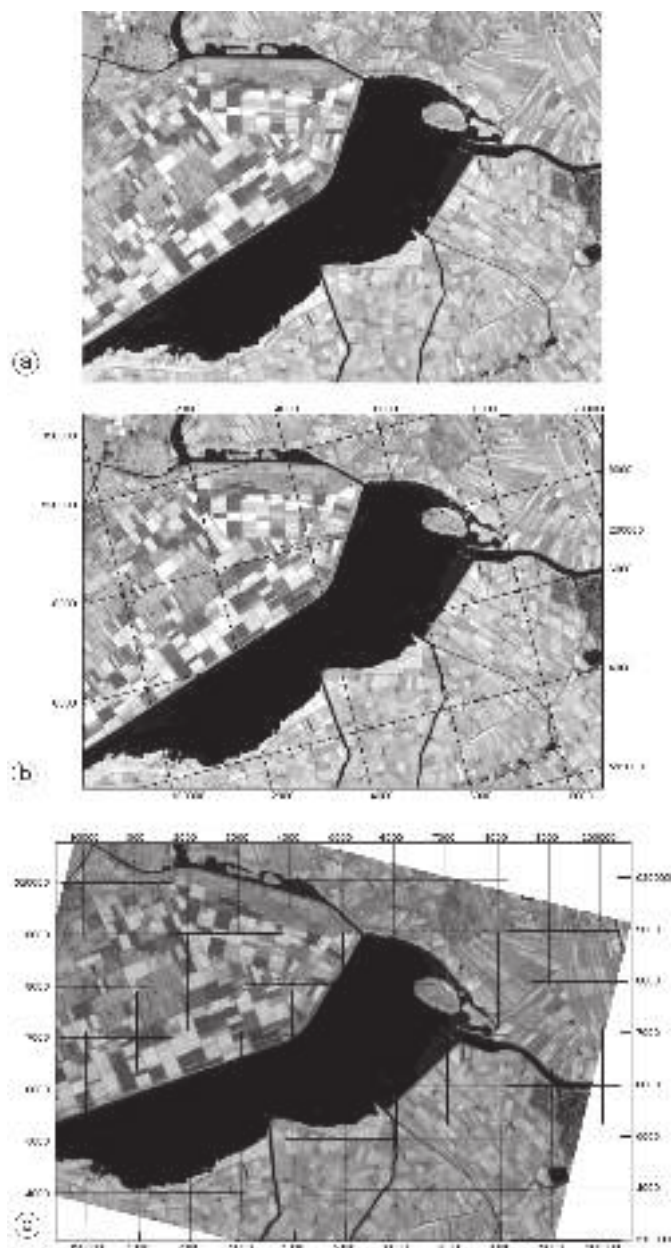
Figure 5.27
Original (top), georeferenced
(middle) and geocoded
(bottom) image of a part of
Flevoland.

stored in the same map projection. The effect of georeferencing and geocoding is illustrated by Figure 5.27. Distinguishing georeferencing and geocoding is conceptually useful, but to the casual software user it is often not apparent whether only georeferencing has been applied in certain image manipulations or also geocoding.

Geocoding is georeferencing with subsequent *resampling* of the image raster. This means that a new image raster is defined along the $x-$ and y-axes of the selected map projection. The geocoding process comprises three main steps: (1) selection of a new grid spacing; (2) projection (using the transformation parameters) of each new raster element onto the original image; and (3) determination and storage of a DN for the

new pixel.

Figure 5.28 shows four transformation types that are frequently used in RS. The types shown increase from left to right in complexity and number of parameters required. In a conformal transformation, the image shape, including right angles, are retained. Therefore, only four parameters are needed to describe a shift along the x- and y-axes, a scale change, and the rotation. However, if you want to geocode an image to make it fit with another image or map, a higher-order transformation may be required, such as a projective or polynomial transformation.
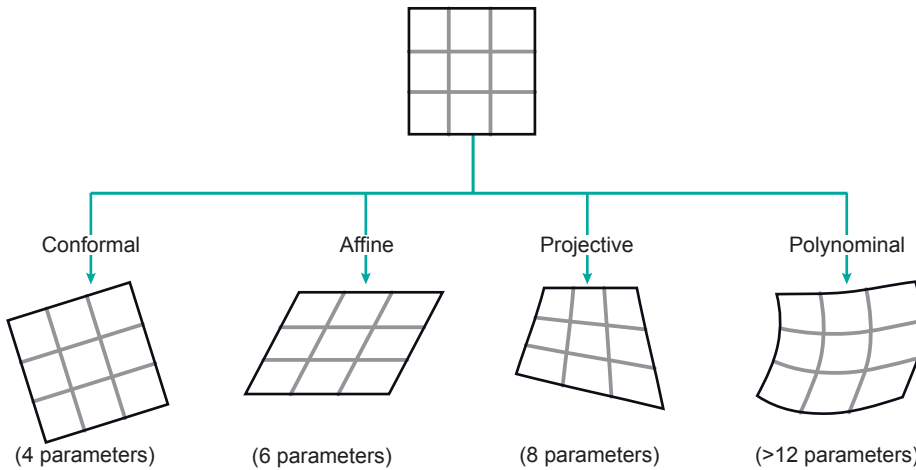
Figure 5.28
Illustration of different image transformation types and the number of required parameters.

Most of the projected raster elements of the new image will not match precisely with raster elements in the original image, as Figure 5.29 illustrates. Since raster data are stored in a regular row–column pattern, we need to calculate DNs for the pixel pattern of the corrected image intended. This calculation is performed by interpolation. This is called *resampling* of the original image.

resampling
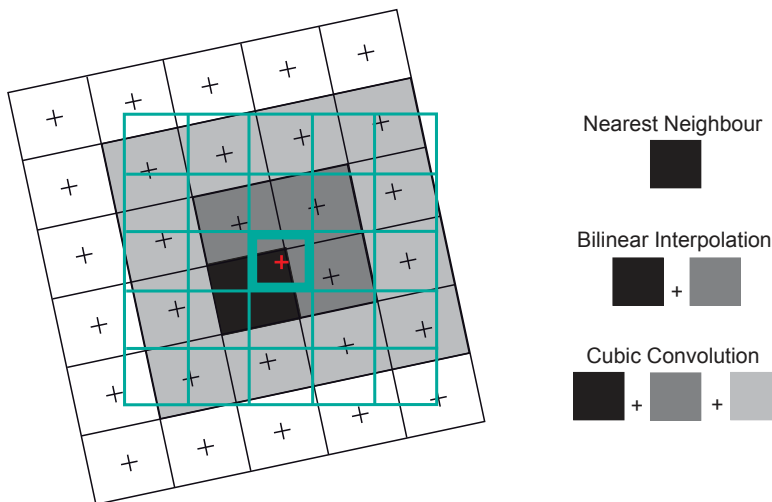
Figure 5.29
Principle of resampling using nearest neighbour, bilinear, and bicubic interpolation.

For resampling, we usually use very simple interpolation methods, the main ones being nearest neighbour, bilinear, and bicubic interpolation (Figure 5.29). Consider the green grid to be the output image to be created. To determine the value of the cen-

tre pixel (bold), in *nearest neighbour interpolation* the value of the nearest original pixel is assigned, i.e. the value of the black pixel in this example. Note that the respective pixel centres, marked by small crosses, are always used for this process. In *bilinear interpolation*, a linear weighted average is calculated for the four nearest pixels in the original image (dark grey and black pixels). In *bicubic interpolation* a cubic weighted average of the values of 16 surrounding pixels (the black and all grey pixels) is calculated. Note that some software uses the terms "bilinear convolution" and "cubic convolution" instead of the terms introduced above.

The choice of the resampling algorithm depends, among other things, on the ratio between input and output pixel size and the intended use of the resampled image. Nearest neighbour resampling can lead to the edges of features being offset in a step-like pattern. However, since the value of the original cell is assigned to a new cell without being changed, all spectral information is retained, which means that the re-sampled image is still useful in applications such as digital image classification (see Section 6.2). The spatial information, on the other hand, may be altered in this process, since some original pixels may be omitted from the output image, or appear twice. Bilinear and bicubic interpolation reduce this effect and lead to smoother images. However, because the values of a number of pixels are averaged, radiometric information is changed (Figure 5.30).

| Original | Nearest Neighbour | Bilinear Interpolation | Bicubic Interpolation |
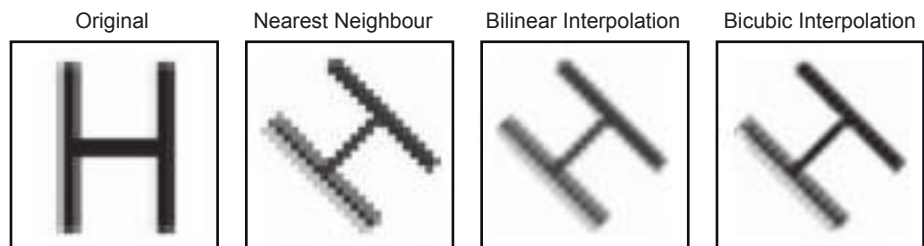
Figure 5.30
The effect of nearest neighbour and bilinear and bicubic resampling of the original data.

5.3.4 Three-dimensional approaches

In mapping terrain, we have to consider its vertical extent (elevation and height) in two types of situations:

- We want 2D geospatial data that describes the horizontal position of terrain features, but the terrain under consideration has large elevation differences. Elevation differences in the scene cause relief displacement in the image. Digitizing in the image without taking into account relief displacement causes errors in computed map coordinates. If the positional errors are larger than would be tolerated by the application (or map specifications), we should not use simple georeferencing and geocoding.

- We want 3D data.

When wishing to map terrain with an increasing degree of refinement, we have to first clarify what we mean by *terrain*. *Terrain* as described by a topographic map has two very different aspects: (1) there are agricultural fields and roads, forests and waterways, buildings and swamps, barren land and lakes; and (2) there is elevation changing with position—at least in most regions on Earth. We refer to land cover, topographic objects, etc. as *terrain features*; we show them on a (2D) map as areas, lines and point symbols. We refer to the shape of the ground surface as *terrain relief*, which

we show on a topographic map by contour lines and/or relief shading. A *contour*

line on a topographic map is a line of constant elevation. Given contour lines, we can determine the elevation at any arbitrary point by interpolating between contour lines.

We could digitize the contour lines of a topographic map. The outcome would be a data set consisting of (X, Y, Z) coordinates of many points. Such a data set is the basis of a *digital terrain relief model* (DTM). We then need a computer program to utilize such a list of coordinate of points of the ground surface, to compute elevation at any horizontal position of interest and derive other terrain relief information such as slope and aspect. The idea of a DTM was developed in the late 1950s at MIT for computerizing highway design. Fifty years later, we use DTMs in all geosciences for a wide range of applications and we have remote sensors specifically built to supply us with data for DTMs (SRTM, SPOT-5 HRS, etc.). One of the applications of a DTM is to accomplish *digital monoplotting* and *orthoimage* production, as outlined later in this subsection.

A *DTM* is a digital representation of terrain relief, i.e. a model of the shape of the ground. We have a variety of sensors at our disposal that can provide us with 3D data: line cameras, frame cameras, laser scanners and microwave radar instruments. They can all produce (X, Y, Z) coordinates of terrain points, but not all the terrain points will be points on the ground surface. Consider a stereo pair of photographs, or a stereo pair from SPOT-5, of a tropical rainforest. Will you be able to see the ground? Coordinates obtained will pertain to points of the terrain relief. Since a model based on such data is not a DTM, we refer to it as digital surface model (DSM). The difference between a DTM and a DSM is illustrated by Figure 5.31; we need to filter DSM data to obtain DTM data.
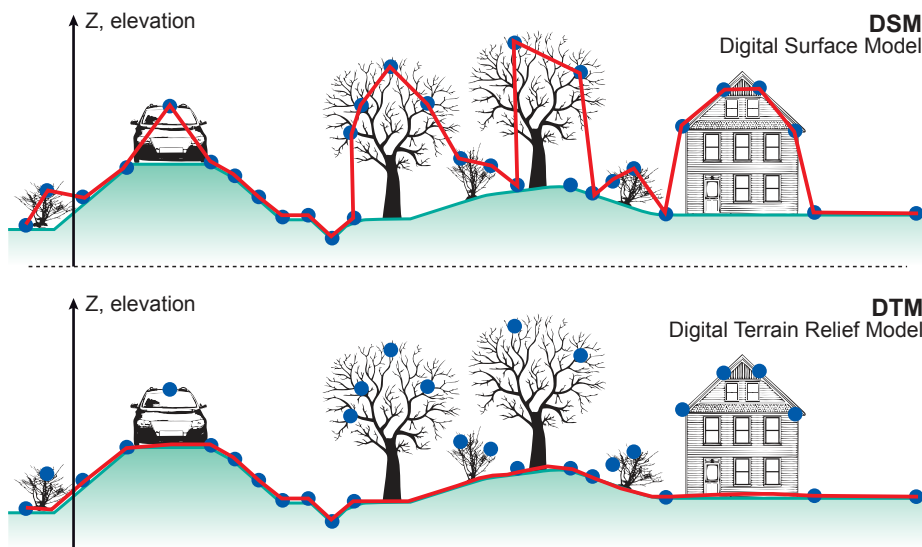
Figure 5.31
The difference between a DTM and a DSM.

In terrain modelling, it is handy to choose the coordinate system such that Z is the variable for elevation. If we model a surface digitally by nothing else than elevation values Z at horizontal positions (X, Y), why not call such a model a *digital elevation model* (DEM)? The term DEM was introduced in the 1970s with the purpose of distinguishing the simplest form of terrain relief modelling from more complex types of digital surface representation. Originally the term DEM was exclusively used for raster representations (thus elevation values given at the intersection nodes of a regular grid). Note that both a DTM and DSM can be a DEM and, moreover, "elevation" would not have to relate to terrain but could relate to some subsurface layer such as groundwater layers, soil layers or the ocean floor. Unfortunately, you will find in the

literature variations in the use of the terms introduced above. This is particularly so for DEM, which is often used carelessly. In this context, it is also worth mentioning the misuse of "topography" as synonym for terrain relief.

5.3.5 Orientation

The purpose of *camera orientation* is to obtain the parameter values for transforming terrain coordinates (X, Y, Z) to image coordinates, and vice versa. In solving the orientation problem we assume that any terrain point and its image lie on a straight line that passes through the projection centre (i.e. the lens). This assumption about the imaging geometry of a camera is called the *collinearity* relationship and it does not take into consideration that atmospheric refraction has the effect of slightly bending light rays. We solve the problem of orienting a single image with respect to the terrain in two steps: *interior orientation* and *exterior orientation*.
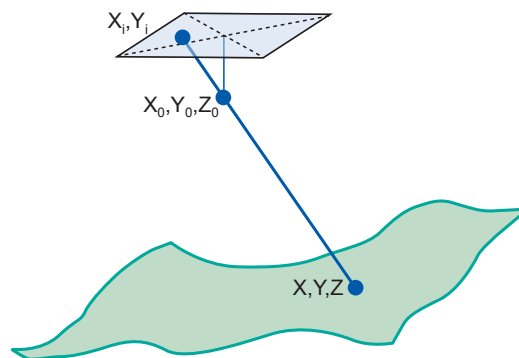
Figure 5.32
Illustration of the collinearity concept, where image point, lens centre and terrain point all lie on one straight line.

Orienting a single image as obtained by a line or frame camera

Interior orientation determines the position of the projection centre with respect to the image. The problem to solve is different for digital cameras and film cameras. In the case of a digital camera (line or frame camera, space-borne or aerial camera), the position of the projection centre with respect to the CCD array does not change from image to image, unless there are extreme temperature or pressure changes. For standard applications, we can assume that the position of the projection centre does not change when defined in the row–column system of the digital image. Two parameters are needed, the *principal distance* and the *principal point* (Figure 5.33); they are both determined by camera calibration. The *principal distance* is the mathematical abstraction of the focal length (which is a physical property of a lens). The *principal point* is the point of intersection of the perpendicular from the projection centre point with the image plane.

interior orientation

principal distance

principal point

The camera calibration report states the row (and column) number of the principal point and the principal distance. When using digital photogrammetric software for working with images from digital aerial cameras, the user only has to enter the principal distance, the position of the principal point and the pixel size to define the interior orientation. In the case of a film camera, the position of the principal point is only fixed in the camera and determined by camera calibration with respect to the fiducial marks. When scanning photographs, the principal point will have different positions in the row–column system of each image, because each image will be placed slightly differently in the scanner. Therefore, you have to measure for every image the imaged fiducial marks and calculate the transformation onto the fiducial marks as given by the calibration report. Digital photogrammetric software can then relate row–column

coordinates of any image point to image coordinates (x, y) at the time of exposure (Figure 5.33).
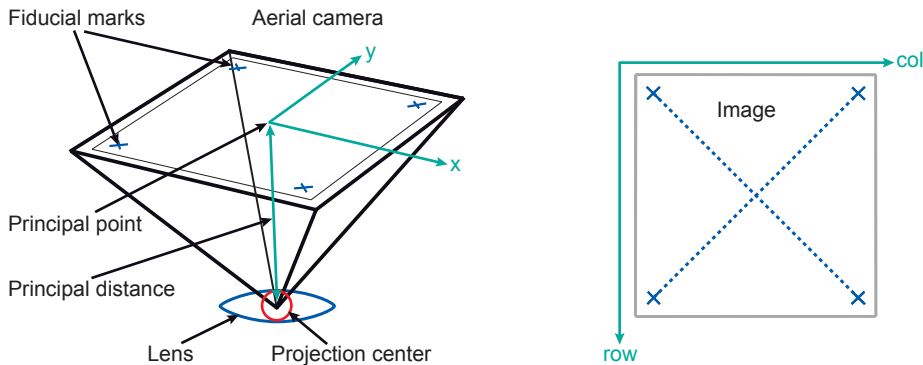
Figure 5.33
Inner geometry of a camera and the associated image.

The *exterior orientation* determines the position of the projection centre with respect to the terrain and also the attitude of the sensor. Frame cameras (film as well as digital ones) acquire the entire image at once, thus three coordinates for the projection centre (X, Y, Z) and three angles for the attitude of the sensor (angles relating to roll, pitch, and yaw) are sufficient to define the exterior orientation of an entire image. Line cameras yield images for which each image line has its own projection centre and the sensor attitude may change from one line to the next. Satellites have a very smooth motion and airborne line cameras are mounted on a stabilized platform so that any attitude change is gradual and small. Mathematically, we can model the variation of an attitude angle through a scene with a low-degree polynomial.

exterior orientation

sensor position and attitude

Images from modern, high-resolution line cameras on satellites come with *rational polynomial coefficients* (RPCs). The rational polynomials define by good approximation the relationship between the image coordinates of an entire frame (in terms of row–column pixel positions) and terrain coordinates. The nice thing is that RPCs are understood by RS software such as ERDAS and that they take care of interior and exterior orientation. For cases in which RPCs are not given or are considered not accurate enough, the exterior orientation needs to be solved in one of the following ways:

RPC

- *Indirect camera orientation*: identify GCPs in the image and measure the row and column coordinates; acquire (X, Y, Z) coordinates for these points, e.g. by GPS or a sufficiently accurate topographic map; use adequate software to calculate the exterior orientation parameters, after having completed the interior orientation.

- *Direct camera orientation*: during image acquisition, make use of GPS and IMU recordings by employing digital photogrammetric software.

- *Integrated camera orientation*, which is a combination of (a) and (b).

For high resolution satellite images such as Ikonos or QuickBird, adding one GCP can already considerably improve the exterior orientation as defined by the RPC. For Cartosat images, it is advisable to improve the exterior orientation by at least five GCPs. For orienting a frame camera image you need at least three GCPs (unless you also have GPS and IMU data). After orientation, you can use the terrain coordinates of any reference point and calculate its position in the image. The differences between measured and calculated image coordinates (the residuals) allow you to estimate the accuracy of orientation. As you may guess, advanced camera/sensor/image orientation is a topic for further study.

Orienting a stereo-image pair obtained by a line or frame camera

The standard procedure is to individually orient each image. In the case of images originating from a frame camera, we can, however, readily make use of the fact that both images partly cover the same area. Instead of doing two independent exterior orientations, we can better first do a *relative orientation* of the two images, followed by an *absolute orientation* of the pair to the terrain coordinate system. The *relative orientation* will cause the imaging rays of corresponding points to intersect more accurately than if you orient one image without the knowledge of the other. You do not have to measure points with known terrain coordinates to solve the relative orientation problem; you only need to measure image coordinates of corresponding points in the two images (after the individual interior orientations have been established). Measuring of corresponding points can even be done automatically (by *image matching*, see below). For absolute orientation, at least three GCPs are needed. The idea of splitting up the exterior orientation into a relative orientation and an absolute orientation is also used for orienting a whole block of overlapping images, not just two. The advantage is that we need a only a few GCPs (which are usually expensive to acquire) and we can still obtain accurate transformation parameters for each image. The method is known as *aerial triangulation*.

relative orientation

absolute orientation

5.3.6 Monoplotting

Suppose you need to derive accurate planimetric coordinates of features expressed in a specific map projection from a single aerial photograph. For flat terrain, this can be achieved using a vertical photograph and a georeferencing approach. Recall from the earlier discussion on relief displacement (Figure 5.24) how elevation differences lead to distortions in the image, preventing the use of such data for direct measurements. Therefore, if there is significant terrain relief, the resulting relief displacement has to be corrected. The method of *monoplotting* was developed (with major research input from ITC) for just this purpose.

Aerial photo   Ground control points   Digital terrain model

(x,y)

Processing

(X,Y,Z)
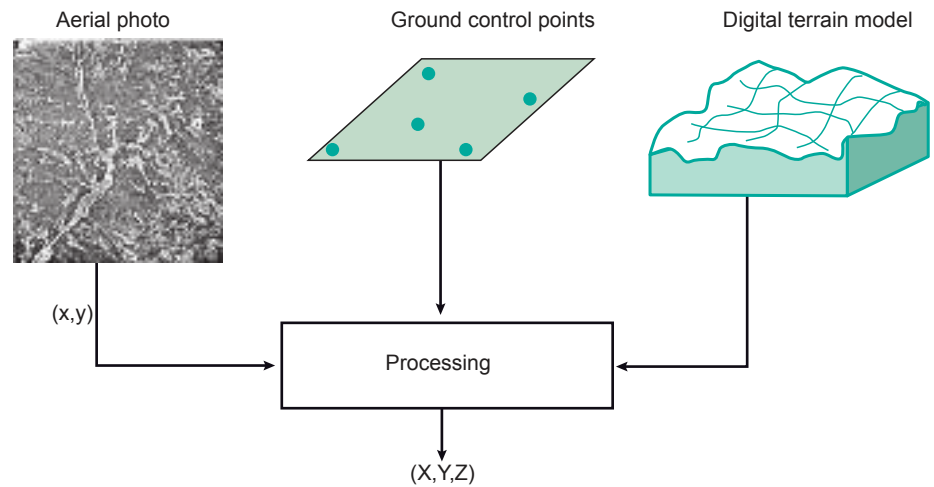
Figure 5.34
The process of digital monoplotting enables accurate determination of terrain coordinates from a single aerial photograph.

Monoplotting is based on the reconstruction of the position of the camera at the moment of image exposure with respect to the GCPs, i.e. the terrain. This is achieved by identifying several (at least four) GCPs for which both the photo and map coordinates are known. Information about the terrain relief is supplied by a DTM of adequate accuracy. The DTM should be given in the required map projection system and the elevations should be expressed in an adequate vertical reference system. When digitizing

3D data from a single image

features from the photograph, the computer uses the DTM to calculate the relief displacement for every point and corrects for it (Figure 5.34). A monoplotting approach is possible by using a hardcopy image on a digitizer tablet or by on-screen digitizing on a computer monitor. In the latter case, vector information can be superimposed on the image to update the changed features. Note that monoplotting is a (real-time) correction procedure and does not yield a new image, i.e. no resampling is carried out.

5.3.7 Orthoimage production

Monoplotting can be considered a georeferencing procedure that incorporates corrections for relief displacement without involving any resampling. For some applications, however, it is useful to actually correct the photograph or RS image, taking into account the effect of terrain relief. In such cases, the image should be transformed and resampled (making use of a DTM) into a product with the geometric properties of a specific map projection. Such an image is called an *orthophoto*.

geocoding an image of rough terrain

The production of orthophotos is quite similar to the process of monoplotting. Consider a scanned aerial photograph. First, the photo is oriented using ground control points. The terrain elevation differences are modelled by a DTM. The computer then calculates the position in the original photo for each output pixel. Using one of the resampling algorithms, the output value is determined and stored in the required raster. The result is geometrically equivalent to a map, i.e. direct distance or area measurements on the orthoimage can be carried out.

5.3.8 Stereo restitution

After relative orientation of a stereo pair, we can exploit the 3D impression gained from the stereo model to make measurements in 3D. The measurements made in a stereo model make use of a phenomenon known as *parallax* (Figure 5.35). Parallax refers to the fact that an object photographed from different camera locations (e.g. from a moving aircraft) has different relative positions in the two images. In other words, there is an apparent displacement of an object as it is observed from different locations. Figure 5.35 illustrates that points at two different elevations, regardless of whether it is the top and bottom of a hill or of a building, experience a relative shift.

3D data from a stereo pair

Parallax

P_a

P_b
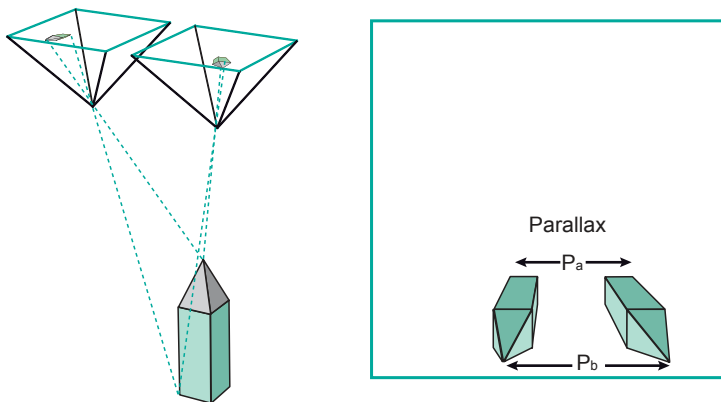
Figure 5.35
The same building is observed from two different positions. Because of the height of the building, the positions of the building top and base relative to the photo centres are different. This difference (parallax) can be used to calculate its height.

image matching

The measurement of the difference in position is basic input for elevation calculations. We could use stereo restitution to measure (X, Y, Z) coordinates of many points, in this way obtaining data for a DTM. This is both a boring and an error-prone process. Digital photogrammetric software can do this job automatically with a high degree of success—after some 30 years of research and development—using *image matching*. Manual measurements are then only needed as a supplement for difficult areas. The main purpose of stereo restitution is to collect 3D vector data of objects. Unlike mono-plotting, elevation is measured directly and not interpolated from a DTM, hence the coordinate accuracy can be higher. Another main advantage of stereo restitution is the better image interpretability obtained, because one can see and interpret in 3D.

analogue,
analytical,
digital systems

A stereo model enables parallax measurement using a special 3D cursor. If the stereo model is appropriately oriented, the parallax measurements yield (X, Y, Z) coordinates. *Analogue* systems use hardcopy images and perform the computation by mechanical, optical or electrical means. *Analytical* systems also use hardcopy images, but do the computation digitally, while in modern *digital* systems, both the images and the computation are digital. By using digital instruments, we cannot only make a few spot elevation measurements, but also generate automatically a complete DSM for the overlapping part of the two images. Recall, however, that reliable elevation values can only be extracted if the orientation steps were carried out accurately, using reliable ground control points.

Chapter 6

Image analysis

Wan Bakx
Lucas Janssen
Ernst Schetselaar
Klaus Tempfli
Valentyn Tolpekin
Eduard Westinga

6.1 Visual image interpretation

6.1.1 Introduction

How to extract information from images? In general, methods for extracting information from remote sensing images can be subdivided into two groups:

- Information extraction based on visual image interpretation. Typical examples of this approach are visual interpretation methods for land use or soil mapping. Acquisition of data from aerial photographs for topographic mapping is also based on visual interpretation.

- Information extraction based on semi-automatic processing by computer. Examples of this approach include automatic generation of DTMs, digital image classification and calculation of surface parameters.

The most intuitive way of extracting information from remote sensing images is by visual image interpretation, which is based on our ability to relate colours and patterns in an image to real world features. Chapter 5 explains the different methods used to visualize remote sensing data. We can interpret images displayed on a computer monitor or printed images, but how to convey our findings to somebody else? In everyday life we often do this verbally, but for thousands of years we have also been doing it by mapping. We used to overlay a transparency on a photograph and trace over the outline of areas that we recognized as having characteristics we were interested in. By doing so for all features of interest in a scene, we obtained a map. The digital variant of this approach is to digitize—either on-screen, or using a digitizer tablet if we only have a hardcopy image—points, lines and areas and label these geometric entities to convey thematic attributes. This way we obtain a map of, for example, all vineyards

mapping

in a certain area and the roads and tracks leading to them. Instead of interpreting and digitizing from a single image, we can also use a stereo-image pair. The interpretation process is the same, although we do need special devices for stereoscopic display and viewing, as well as equipment that allows us to measure properly in a stereogram.

Visual image interpretation is not as easy as it may seem at first glace; it requires training. Yet our eye–brain system is quite capable of doing the job. Visual interpretation is, in fact, an extremely complex process, as was discovered when we tried to let computers do image interpretation. Research on *image understanding* has helped us to conceptualize human vision and interpretation, and progress in this area continues to be made.

Subsection 6.1.2 explains the basics of how we recognize features and objects in images. Visual image interpretation is used to produce geospatial data in all of ITC's fields of interest: urban mapping, soil mapping, geomorphological mapping, forest mapping, natural vegetation mapping, cadastral mapping, land use mapping, and many others. Actual image interpretation is application specific, although it does follow a standard approach. Subsection 6.1.3 describes this general, practical approach. Aspects of assessing the quality of the outcome of an interpretation are treated in Subsection 6.1.4.

6.1.2 Interpretation fundamentals

Human vision

Human perception of colour is explained in Chapter 5. *Human vision* goes a step beyond the perception of colour: it deals with the ability of a person to draw conclusions from visual observations. When analysing an image, typically you find yourself somewhere between the following two processes: direct and *spontaneous recognition*; and *logical inference*, using clues to draw conclusions by a process of reasoning.

Spontaneous recognition refers to the ability of an interpreter to identify objects or features at first glance. Consider for a moment Figure 6.1. Agronomists would immediately recognize the pivot irrigation systems from their circular shape. They are able to do so because of earlier (professional) experience. Similarly, most people can directly relate what they see on an aerial photo to the terrain features of the place where they live (because of "scene knowledge"). The statement made by people that are shown an aerial photograph of their living environment for the first time, "I see because I know", is rooted in spontaneous recognition.

spontaneous recognition

As the term states, *Logical inference* means that the interpreter applies reasoning. In the reasoning, the interpreter uses acquired professional knowledge and experience. Logical inference is, for example, concluding that a rectangular shape is a swimming pool because of its location in a backyard garden near to a house. Sometimes logical inference alone is insufficient to interpret images; then field observations are required (see Subsection 6.1.3). Consider the aerial photograph in Figure 6.2. Are you able to interpret the material and function of the white mushroom-like objects? A field visit would be required for most of us to relate the different features to elements of a house or settlement.

logical inference

Interpretation elements

We need a set of terms to express the characteristics of an image that we can use when interpreting the image. These characteristics are called *interpretation elements* and are used, for example, to define *interpretation keys*, which provide guidelines on how to recognize certain objects.

The following seven interpretation elements can be distinguished: tone/hue, texture,

pattern, shape, size, height/elevation, and location/association.

- *Tone* is defined as the relative brightness in a B&W. *Hue* refers to the colour as defined in IHS colour space. Tonal variations are an important interpretation element. The tonal expression of objects in an image is directly related to the amount of light (or other forms of EM radiation) reflected (or emitted) from the surface. Different types of rock, soil or vegetation are most likely have different tones. Variations in moisture conditions are also reflected as tonal differences in an image: increasing moisture content gives darker grey tones. Variations in hue are primarily related to the spectral characteristics of the imaged terrain and also to the bands selected for visualization (see Chapter 5). The advantage of hue over tone is that the human eye has a much larger sensitivity for variations in colour (approximately 10,000 colours) than tone (approximately 200 grey levels).

- *Texture* relates to the frequency of tonal change. Texture may be described by terms such as coarse or fine, smooth or rough, even or uneven, mottled, speckled, granular, linear, woolly, etc. Texture can often be related to terrain surface roughness. Texture is strongly related to the spatial resolution of the sensor used. A pattern on a large-scale image may show as texture on a small-scale image of the same scene.

- *Pattern* refers to the spatial arrangement of objects and implies the characteristic repetition of certain forms or relationships. Pattern can be described by terms such as concentric, radial and checkerboard. Some land uses have specific and characteristic patterns when observed from the air or space. Different types of irrigation may spring to mind, or different types of housing on an urban fringe.

Figure 6.2
Mud huts of Labbezanga, near the Niger river (Photo by Georg Gerster, 1972).

Other typical examples include hydrological systems (a river and its tributaries) and patterns related to erosion.

- *Shape* or form characterizes many objects visible in an image. Both the two-dimensional projection of an object, as shown on a map, and the height of an object influence the shape of its image. The shape of objects often helps us to identify them (built-up areas, roads and railroads, agricultural fields, etc.).

- *Size* of objects can be considered in a relative or absolute sense. The width of a road can be estimated, for example, by comparing it to the size of the cars using it, which is generally known. Subsequently, the width determines the road type, e.g. primary road, secondary road, and so on.

- *Height* differences are important for distinguishing between different vegetation types, building types, etc. Elevation differences provide us with clues in geomorphological mapping. We need a stereogram and stereoscopic viewing to observe height and elevation. Stereoscopic viewing facilitates interpretation of both natural and man-made features.

- *Location/association* refers to situation of an object in the terrain or in relation to surroundings. A forest in the mountains is different from a forest close to the sea or one near a meandering river. A large building at the end of a number of converging railroads is likely to be a railway station—we would not expect a hospital at such a location.

With these seven interpretation elements, you may have noticed a relation with the spatial extent of the feature to which they relate. Tone or hue can be defined for a single pixel; texture is defined for a group of adjacent pixels, not for a single pixel. The other interpretation elements relate to individual objects or a combination of objects. The simultaneous and often intuitive use of all these elements is the strength of visual image interpretation. In standard digital image classification (Section 6.2) only hue is utilized, which explains the limitations of automated methods compared to visual image interpretation.

6.1.3 Mapping

Interpretation

The assumption in mapping with the help of remote sensing images is that areas that look homogeneous in the image will have similar features on the ground. The interpretation process consists of delineating areas that internally appear similar and at the same time different from other areas. Making an interpretation from only one aerial photograph or a small part of an image from a space-borne sensor seems quite simple. You have the overview of the entire area at all times and can easily compare one unit to another and decide if they are the same or different. Working with many photographs and also with several people will, in contrast, require a clear definition of the units to be delineated.

Definition of units is based on what can be observed in the image. Different interpretation units can be described according the interpretation elements. After establishing what the features are on the ground, *'interpretation elements* can be constructed, from which an interpretation of features can be made. These features are again described in terms of interpretation elements. If knowledge of the area is lacking (not yet available), you could also begin your interpretation based only on interpretation elements (Figure 6.3). After fieldwork, it will become clear what the units actually represent on the ground.

interpretation key

1 grey dense drainage pattern
2 grey + some black dots
3 black course texture
4 grey + dense black dots
5 grey
6 white
7 dark grey
8 black smooth texture
9 grey fine texture
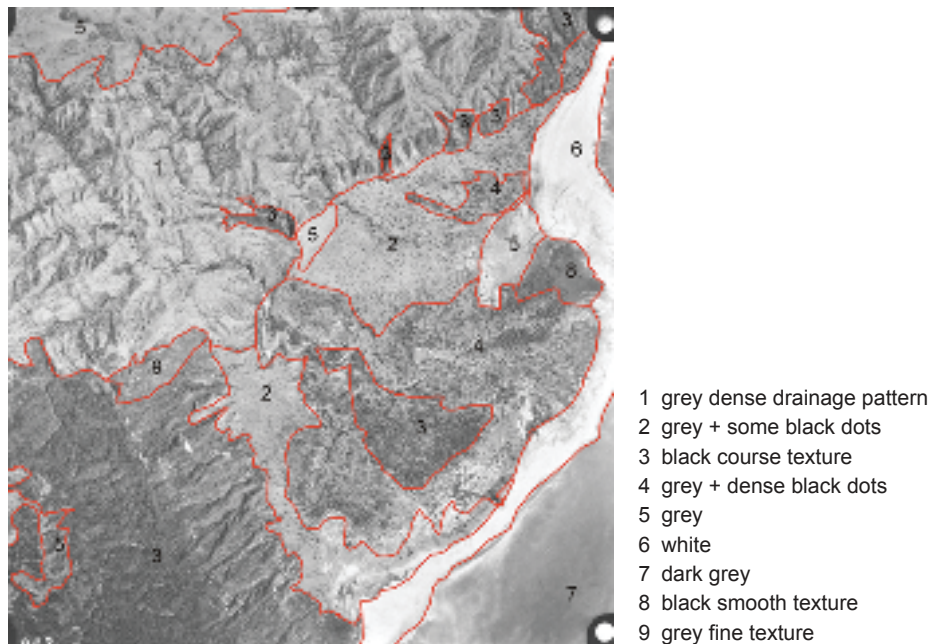
Figure 6.3
Example of an interpretation
of Manyara, Tanzania.

Prior to the delineation of the units, a legend is constructed based on interpretation elements. The legend can be presented in the form of a table in which each element type is represented by a column. Table 6.1 presents a fictitious example of a legend. In this legend, the "unit number" represents an as yet unknown feature type; the corresponding row elements will be used to identify that feature type.

interpretation legend

When preparing a legend you need to consider that distinguishing units can be based on a difference in one element only or on differences of several elements. For example, consider Unit 1 in Table 6.1: its tone is black and all other units have a grey or white tone. In this case there is no need to define all the other elements for Unit 1. In the

Table 6.1
Fictitious example of an
interpretation legend.

Unit	Tone	Texture	Shape	Size	Height	Location
1	black					
2	grey	smooth				
3	grey	rough			high	mountains
4	grey	rough			low	
5	grey	rough			high	sea + river
6	grey		field			
	white		line			
7	white		field			
	grey		line			
8	grey + black		field	square		
9	grey + black		field	rectangle 5 × 5		
10	grey + black		field	rectangle 20 × 20		

example of Figure 6.3, some units are different in texture. There are areas with smooth and rough texture. The rough texture areas are further differentiated according to height. Furthermore, rough, high areas are differentiated, depending on location in the mountains or along rivers or near the sea.

When delineating areas by hand, there is a limit to what can still be drawn. In practice, polygons smaller than 5 mm × 5 mm should not be drawn. This is called the smallest allowable unit. The scale of the image(s) used therefore limits the interpretation cell on the ground. When delineating areas by digitizing on-screen, one could zoom in—in principle to a monitor dot. However, you need to define the maximum scale at which the given remote sensing data are still reliable and then calculate the smallest allowable unit.

interpretation cell

In some cases an area may consist of two or three different types of too-small areas. Then, individual polygons for each small area cannot be drawn, even though at a larger scale the individual features could be mapped. The solution in such a case is to combine these areas to form a complex unit. The different features of such a complex can be described separately. In Table 6.1, Unit 6 and Unit 7 are two different complex units: in Unit 6 there are two features, namely grey fields and white lines, while in Unit 7 there are white fields and grey lines.

complex unit

Fieldwork

Maps and inventories should reflect what is actually on the ground. Field visits should, therefore, be made to observe what is there in reality. Field visits for ground observation are time-consuming and usually costly. Making observations everywhere in the entire area to be mapped is likely to take too much time. For reasons of efficiency, remote sensing data are used to extrapolate the results of a limited number of observations over the entire area being studied.

The selection of sample locations is a crucial step for cost-effective mapping. We can use the RS images to stratify the area. To do this, we make an preliminary interpretation of the area to be mapped based on its interpretation elements. The interpretation units are the strata to be sampled. For all strata, an equal number of samples are taken; this is known as *stratified sampling*. We can select the samples in such away that they are representative for the interpretation elements of that unit (strata). This is called *stratified representative sampling*. Stratified representative sampling is a very time-efficient and cost-effective method as compared to random or systematic sampling ([132]; [41]). If an interpretation unit occupies a very small area, many samples would be needed for random or systematic sampling, to make sure that small units are also sampled. When applying the stratified sampling approach, far fewer samples

stratified sampling

are needed.

Stratified representative sampling can only be applied if the data to be mapped are qualitative (i.e. nominal or ordinal). For mapping of quantitative data (i.e. interval or ratio data), *unbiased sampling* strategies (i.e. random or systematic sampling) should be applied to allow statistical analysis. Biomass measurements are an example of quantitative data. Then the entire area needs to be sampled and no prior interpretation is needed for the sampling strategy. Both stratified and unbiased sampling strategies will be used if quantitative data of certain strata are not required. For instance, we use *stratified random sampling* of grass biomass for livestock management if in the strata forest, water and urban areas no biomass measurements are needed. We do so to limit time-consuming, unbiased sampling procedures.

unbiased sampling

During fieldwork, the locations of boundaries of the interpretation are also verified. In addition, data is gathered about areas or features that cannot be derived from remote sensing images.

Analysing field data and map preparation

From the correlation between collected field data and the interpretation, the entire area can be mapped in terms of what is on the ground. If there is a good correlation, only recoding and renaming of the units will be required. If the correlation is poor, however, a complete re-interpretation might be needed, after carefully restudying the legend in terms of interpretation elements. For producing the final map, all aspects of map design and cartographic finishing should be observed; this is treated in Section 10.1).

6.1.4   Quality aspects

The quality of the result of image interpretation depends on three factors: the interpreter, the images used, and the guidelines provided.

- Professional experience, particularly experience with image interpretation, determines the skill of a photo-interpreter. A professional background is required: a geological interpretation, for example, can only be made by a geologist, since they are best able to related image features to geological phenomena. Local knowledge, derived by field visits, is needed to facilitate interpretation.

trained interpreter

- The images used limit the phenomena that can be studied, both in a thematic and geometric sense. One cannot, for example, generate a reliable database on tertiary road systems using data from low resolution multispectral scanners. On the other hand, B&W aerial photos contain limited information about agricultural crops.

adequate images

- The quality of the interpretation guidelines has a great deal of influence. Consider, for example, a project in which a group of persons is to carry out a mapping project. Ambiguous guidelines will prevent consistent mapping, for which a seamless database of consistent quality is required, despite individual input.

clear guidelines

Especially in large projects and monitoring programmes, all three points just listed play an important role in ensuring the replicability of the work. *Replicability* refers to the degree of correspondence of results obtained by different persons for the same area or by the same person for the same area at different times. Replicability does not provide information on the accuracy (the relation with the real world) but it does give an indication of the quality of the class definition (crisp or ambiguous) and the instructions and methods used. Figure 6.4 and Figure 6.5 provide two examples of how this works. Figure 6.4 gives two interpretation results for the same area. Note

replicability

that both results differ in terms of the total number of objects (map units) and in terms of (line) generalization. Figure 6.5 compares 13 individual geomorphological interpretations. Similarly to Figure 6.4, large differences occur along the boundaries. In addition to this, you could also conclude that for some objects (map units) there was no agreement on the thematic attribute.

Figure 6.4
Two interpretation results derived by two photo-interpreters analysing the same image. Note the overall differences, but also differences in generalization of the lines. (From [73].)
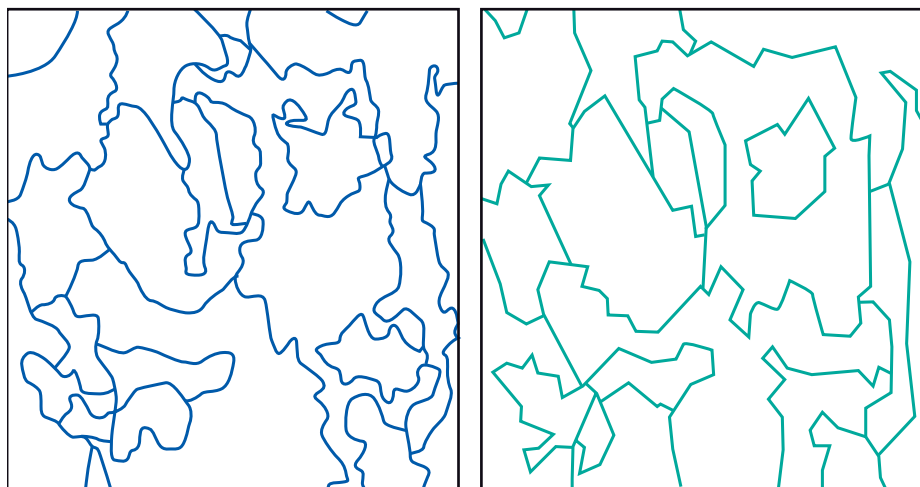
Figure 6.5
Comparison of 13 interpretations of the same image. The grey value represents the degree of correspondence: white indicates agreement of all 13 interpreters; black indicates that all 13 interpreters disagreed on the thematic class for that location. (From [73].)
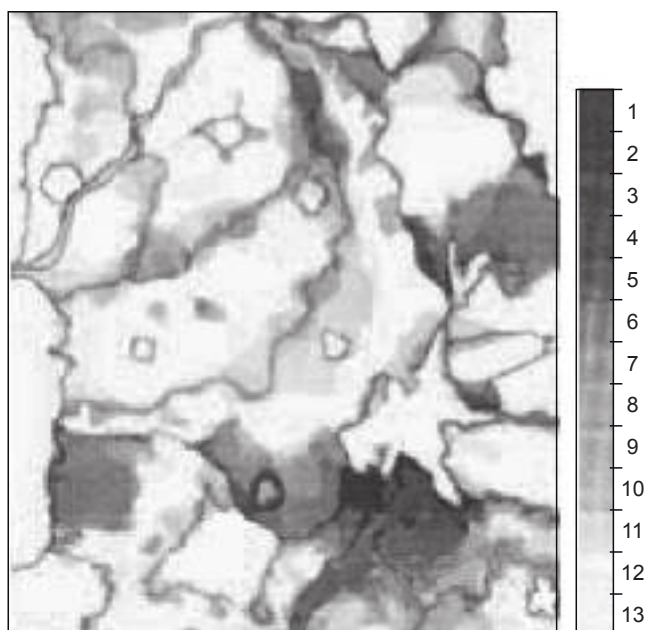
6.2 Digital image classification

Introduction

The process of visual image interpretation has been explained in Section 6.1. In this process, human vision plays a crucial role in extracting information from images. Although computers may be used for visualization and digitization, the interpretation itself is done by the human operator.

This section introduces *digital image classification*. In this process, the human operator instructs the computer to perform an interpretation according to certain conditions, which are defined by the operator. Image classification is one of the techniques in the domain of digital image interpretation. Other techniques include automatic object recognition (for example, road detection) and scene reconstruction (for example, generation of 3D object models). Image classification is the most commonly applied technique in ITC's fields of interest.

digital image interpretation

Image classification is applied in many regional-scale projects. In Asia, the Asian Association of Remote Sensing (AARS) is generating various sets of land cover data based on supervised and unsupervised classification of multispectral satellite data. In the Africover project (an FAO initiative), techniques for digital image classification are being used to establish a pan-African land cover data set. The European Commission requires national governments to verify the claims of farmers related to crop subsidies. To meet these requirements, national governments employ companies to make an intial inventory, using image classification techniques, which is followed later by field checks.

regional scale projects

Image classification is based on the different spectral characteristics of different materials, as introduced in Section 2.5. Here, in section 6.2, the focus is on the classification of multispectral data. Subsection 6.2.1 explains the concepts of image space and feature space, where image classification (Subsection 6.2.2) takes place. Section 6.2.3 gives an overview of the classification process, the steps involved and the choices to be made. The results of image classifications need to be validated to assess their accuracy, the topic of Subsection 6.2.4). Subsection 6.2.5 discusses the problems of standard classification and introduces object-oriented classification.

6.2.1 Principles of image classification

Image space

A digital image is a 2D array of pixels. The value of a pixel, i.e. its DN, is in the case of an 8-bit record in the range 0 to 255. Each DN corresponds to the EM radiation reflected or emitted from a ground resolution cell—unless the image has been resampled. The spatial distribution of the DNs defines the image or *image space*. A multispectral sensor records the radiation from a particular GRC in different channels according to its spectral band separation. A sensor recording in three bands (Figure 2.24) yields three pixels with the same row and column tuple (i, j) since they stem from one and the same GRC.

Feature space

When we consider a two-band image, we can say that the two DNs for a GRC are components of a two-dimensional vector $[v_1, v_2]$, the *feature vector* (Figure 6.6). An example of a feature vector is $[13, 55]$, which indicates that the conjugate pixels of band 1 and band 2 have the DNs 13 and 55. This vector can be plotted in a two-dimensional graph.

feature vector

Similarly, we can visualize a three-dimensional feature vector $[v1, v2, v3]$ of a cell in a

three-band image found in a three-dimensional graph. A graph that shows the feature vectors is called a *feature space*, or *feature space plot* or *scatter plot*. Figure 6.6 illustrates how a feature vector (related to one GRC) is plotted in the feature space for two and three bands. Two-dimensional feature-space plots are the most common.

Note that plotting the values is difficult for a four- or more-dimensional case, even though the concept remains the same. A practical solution when dealing with four or more bands is to plot all possible combinations of two bands separately. For four bands, this already yields six combinations: bands 1 and 2, 1 and 3, 1 and 4, bands 2 and 3, 2 and 4, and bands 3 and 4.

scatterplot

Plotting all the feature vectors of a digital image pair yields a 2D scatterplot of many points (Figure 6.7). A 2D scatterplot provides information about pixel value pairs that occur within a two-band image. Note that some combinations will occur more frequently, which can be visualized by using intensity or colour (as introduced in Section 5.1).

Distances and clusters in the feature space

We use distance in the feature space to accomplish classification. Distance in the feature space is measured as *Euclidian distance* in the same units as the DNs (the unit of the axes). In a two-dimensional feature space, the distance between feature vectors $[v_{11}, v_{12}]$ and $[v_{21}, v_{22}]$ can be calculated according to Pythagoras' theorem:

$$d^2 = (v_{21} - v_{11})^2 + (v_{22} - v_{12})^2$$

For the situation shown in Figure 6.8, the distance between $[10, 10]$ and $[40, 30]$ is:

$$d = \sqrt{(40 - 10)^2 + (30 - 10)^2}$$

.
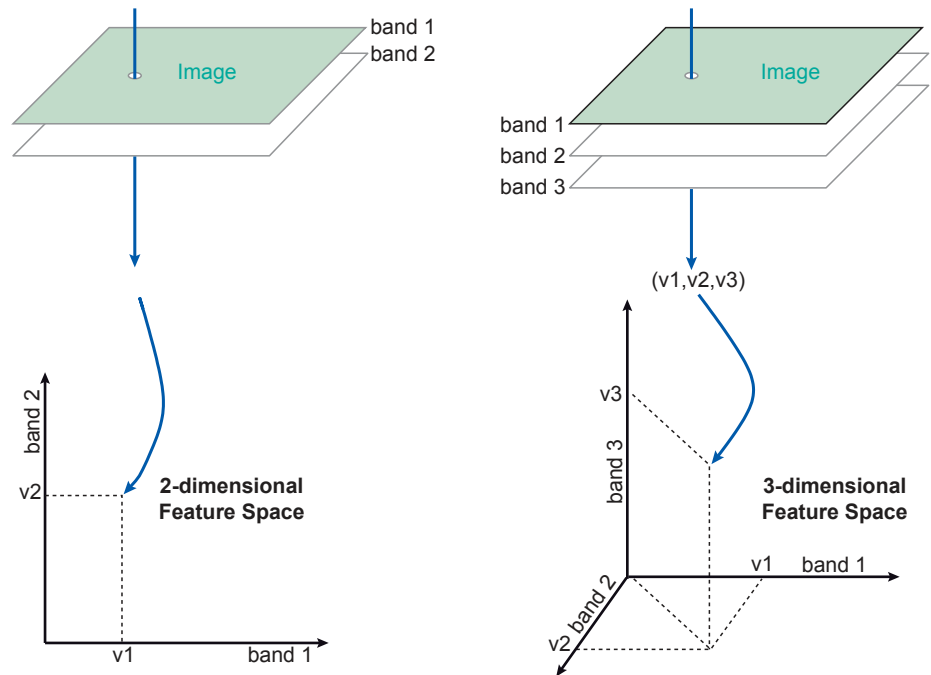
Figure 6.6
Plotting of the pixel values of a GRC in the feature space for a two-band and three-band image.
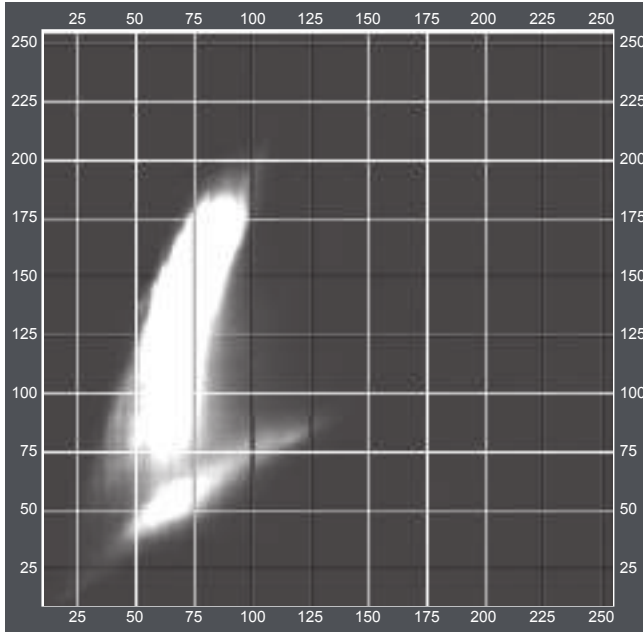
Figure 6.7
Scatterplot of two bands of a
RS image. Note the units
along the x- and y-axes. The
intensity at a point in the
feature space is related to the
number of cells at that point.

For three or more dimensions, the distance is calculated in a similar manner.
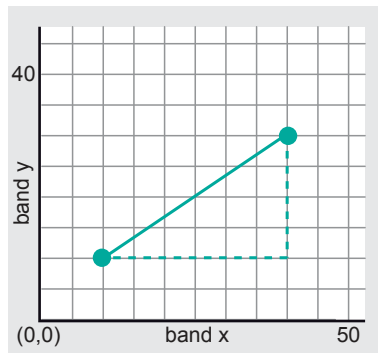
Figure 6.8
Euclidian distance between
the two points is calculated
using Pythagoras' theorem.

6.2.2 Image classification

The scatterplot shown in Figure 6.7 shows the distribution of conjugate pixel values
of an actual two-band image. Figure 6.9 shows a feature space in which the feature
vectors have been plotted of samples of five specific land cover classes (grass, water,
trees, etc.). You can see that the feature vectors of GRCs that are water areas form
a compact cluster. The feature vectors of the other land cover types (classes) are also
clustered. Figure 6.9 illustrates the basic assumption for image classification: a specific
part of the feature space corresponds to a specific class. Once the classes have been
defined in the feature space, each feature vector of a multi-band image can be plotted
and checked against these classes and assigned to the class where it fits best.

cluster

classes

Classes to be distinguished in an image classification need to have different spectral
characteristics. This can, for example, be analysed by comparing spectral reflectance
curves (Section 2.5). This brings us to an important limitation of image classification: if
classes do not have distinct clusters in the feature space, image classification can only

spectral differentiation

give results to a certain level of reliability.
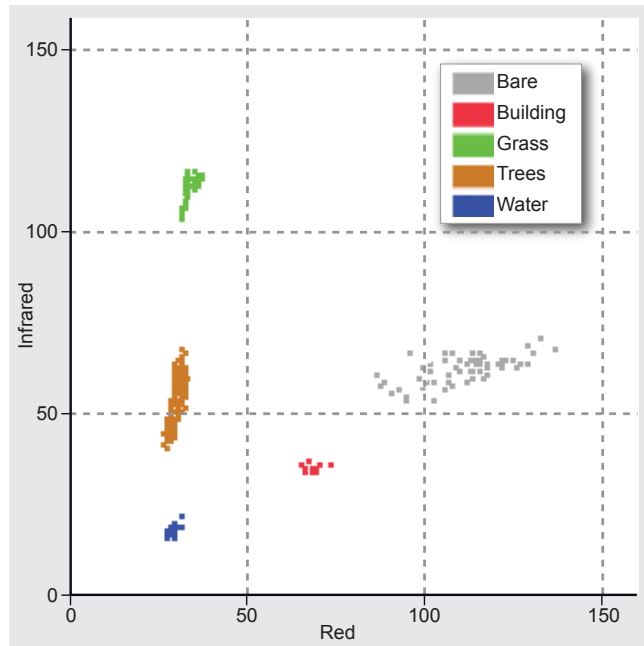
Figure 6.9
Feature space showing the
respective clusters of five
classes; note that each class
occupies a limited area in the
feature space.

The principle of image classification is that a pixel is assigned to a class based on its feature vector, by comparing it to predefined clusters in the feature space. Doing so for all pixels results in a classified image. The crux of image classification is in comparing it to predefined clusters, which requires definition of the clusters and methods for comparison. Definition of the clusters is an interactive process and is carried out during the *training process*. Comparison of the individual pixels with the clusters takes place using *classifier algorithms*. Both of these concepts are explained in the next subsection.

6.2.3 Image classification process

The process of image classification typically involves five steps (Figure 6.10):

data selection

1. Selection and preparation of the RS images. Depending on the land cover types or whatever needs to be classified, the most appropriate sensor, the most appropriate date(s) of acquisition and the most appropriate wavelength bands should be selected.
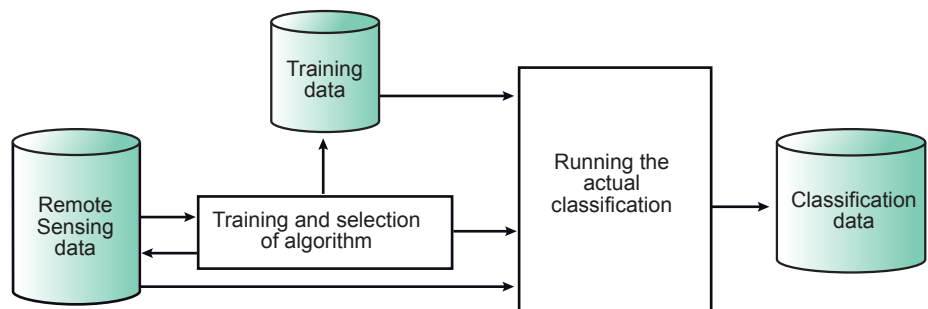
Figure 6.10
The classification process;
the most important
component is the training, in
combination with selection of
the algorithm.

2. Definition of the clusters in the feature space. Here two approaches are possible: *supervised classification* and *unsupervised classification*. In a supervised classification, the operator defines the clusters during the training process; in an unsupervised classification, a clustering algorithm automatically finds and defines the number of clusters in the feature space.

<div align="right">supervised or unsupervised</div>

3. Selection of the classification algorithm. Once the spectral classes have been defined in the feature space, the operator needs to decide on how the pixels (based on their feature vectors) are to be assigned to the classes. The assignment can be based on different criteria.

<div align="right">algorithm selection</div>

4. Running the actual classification. Once the training data have been established and the classifier algorithm selected, the actual classification can be carried out. This means that, based on its DNs, each "multi-band pixel" (cell) in the image is assigned to one of the predefined classes (Figure 6.11).

<div align="right">classifying</div>

5. Validation of the result. Once the classified image has been produced its quality is assessed by comparing it to reference data (ground truth). This requires selection of a sampling technique, generation of an error matrix, and the calculation of error parameters (Subsection 6.2.4).

<div align="right">validating</div>
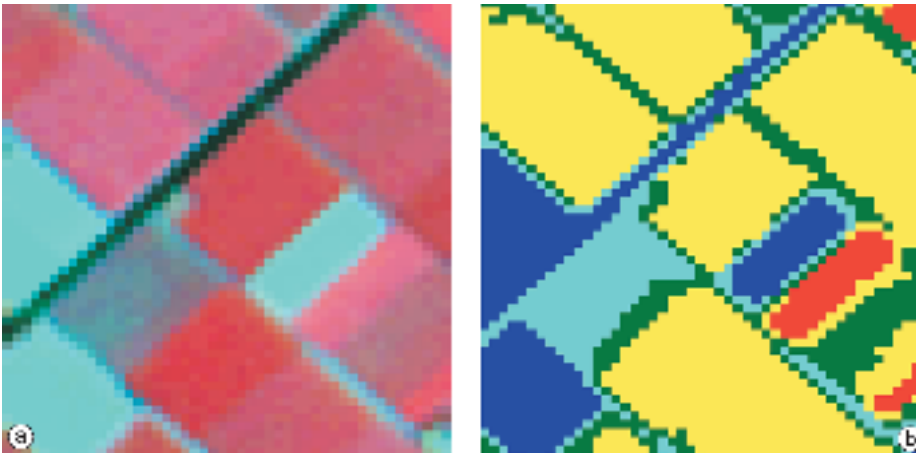
Figure 6.11
The result of classification of a multi-band image (a) is a raster in which each cell is assigned to some thematic class (b).

Each of the steps above are elaborated on in the next subsections. For simplicity and ease of visualization, most examples deal with a two-dimensional situation (two bands), although in principle image classification can be carried out on any n-dimensional data set. Visual image interpretation, however, limits itself to an image that is composed of a maximum of three bands.

Preparation for image classification

Image classification serves a specific goal: converting RS images to thematic data. In the context of a particular application, one is rather more interested in the thematic characteristics of an area (a GRC) than in its reflection values. Thematic characteristics such as land cover, land use, soil type or mineral type can be used for further analysis and input into models. In addition, image classification can also be considered as data reduction: the n multispectral bands result in a single-band image file.

<div align="right">thematic classes</div>

With the particular application in mind, the information classes of interest need to be defined. The possibilities for the classification of land cover types depend on the date

an image was acquired. This not only holds for crops, which have a certain growth cycle, but also for other applications such as snow cover or illumination by the Sun. In some situations, a multi-temporal data set is required. A non-trivial point is that the required images should be available at the required moment. Limited image acquisition and cloud cover may force one to make use of a less-optimal data set.

data date

Before starting to work with the acquired data, a selection of the available spectral bands may be made. Reasons for not using all available bands (for example all seven bands of Landsat-5 TM) lie in the problem of band correlation and, sometimes, in limitations of hardware and software. Band correlation occurs when the spectral reflection is similar for two bands. The correlation between the green and red wavelength bands for vegetation is an example: a low reflectance in green correlates with a low reflectance in red. For classification purposes, correlated bands give redundant information and might disturb the classification process.

selection of bands

Supervised image classification

One of the main steps in image classification is the "partitioning" of the feature space. In supervised classification this is done by an operator who defines the spectral characteristics of the classes by identifying sample areas (training areas). Supervised classification requires that the operator is familiar with the area of interest: the operator needs to know where to find the classes of interest in the scene. This information can be derived from general knowledge of the scene or from dedicated field observations.

scene knowledge

A sample of a specific class, comprising a number of training cells, forms a cluster in the feature space (as portrayed in Figure 6.9). The clusters selected by the operator:

training set

- should form a representative data set for a given class. This means that the variability of a class within the image should be taken into account. Also, in an absolute sense, a minimum number of observations per cluster is required. Although it depends on the classifier algorithm to be used, a useful rule of thumb is $30 \times n$ (n = number of bands) observations.

- should not or only partially overlap with the other clusters, otherwise a reliable separation is not possible. For a specific data set, some classes may have significant spectral overlap, which, in principle, means that these classes cannot be discriminated by image classification. Solutions are to add other spectral bands, and/or add images acquired at other moments.

The resulting clusters can be characterized by simple statistics of the point distributions. These are for one cluster: the vector of mean values of the DNs (for band 1 and band 2) (see Figure 6.14), and the standard deviations of the DNs (for band 1 and band 2) (see Figure 6.15, where the standard deviations are plotted as crosses).

Unsupervised image classification

Supervised classification requires knowledge of the area of interest. If this knowledge is insufficiently available, or the classes of interest have not yet been defined, an unsupervised classification can be made. In an unsupervised classification, clustering algorithms are used to partition the feature space into a number of clusters.

Several methods of unsupervised classification exist, their main purpose being to produce spectral groupings based on certain spectral similarities. In one of the most common approaches, the user has to define the maximum number of clusters in a data set. Based on this, the computer locates arbitrary mean vectors as the centre points of the clusters. Each pixel is then assigned to a cluster by the *minimum distance to cluster centroid* decision rule. Once all the cells have been labelled, recalculation of the cluster

number of classes

Iteration 0

Iteration 1

Iteration 3
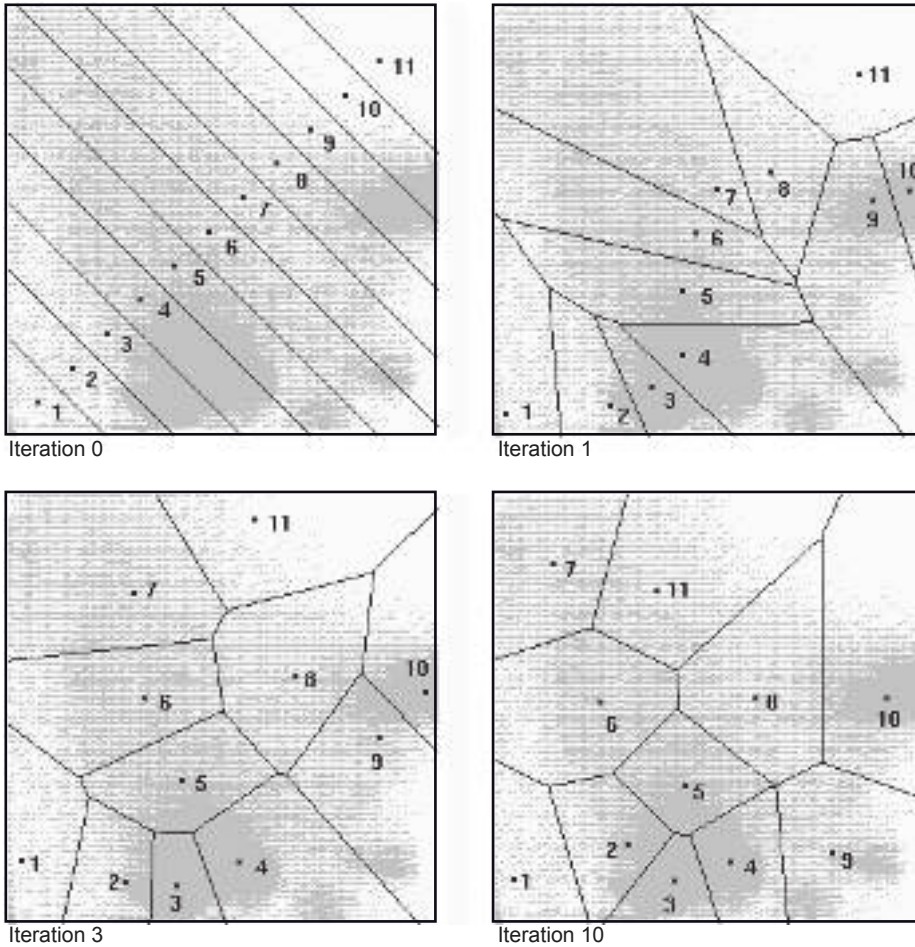
Iteration 10

Figure 6.12
The subsequent results of an
iterative clustering algorithm
on a sample data set.

centre takes place and the process is repeated until the proper cluster centres are found
and the cells are labelled accordingly.

iterative process

The iteration stops when the cluster centres no longer change. However, for any it-
eration, clusters with less than a specified number of cells are eliminated. Once the
clustering is finished, analysis of the closeness or separability of the clusters takes
place by means of inter-cluster distance or divergence measures.

Merging of clusters needs to be done to reduce the number of unnecessary subdivi-
sions in the data set. This is be done using a pre-specified threshold value. The user
has to define the maximum number of clusters/classes, the distance between two clus-
ter centres, the radius of a cluster, and the minimum number of cells as a threshold for
cluster elimination. Analysis of the cluster compactness around its centre point is done
by means of the user-defined standard deviation for each spectral band. If a cluster is
elongated, separation of the cluster will be done perpendicularly to the spectral axis
of elongation.

Analysis of closeness of the clusters is carried out by measuring the distance between
the two cluster centres. If the distance between two cluster centres is less than the pre-
specified threshold, merging of the clusters takes place. The clusters that result after
the last iteration are described by their statistics. Figure 6.12 shows the results of a

clustering algorithm on a data set. Note that the cluster centres coincide with the high density areas in the feature space.

Similarly to the supervised approach, the derived cluster statistics are then used to classify the complete image using a selected classification algorithm.
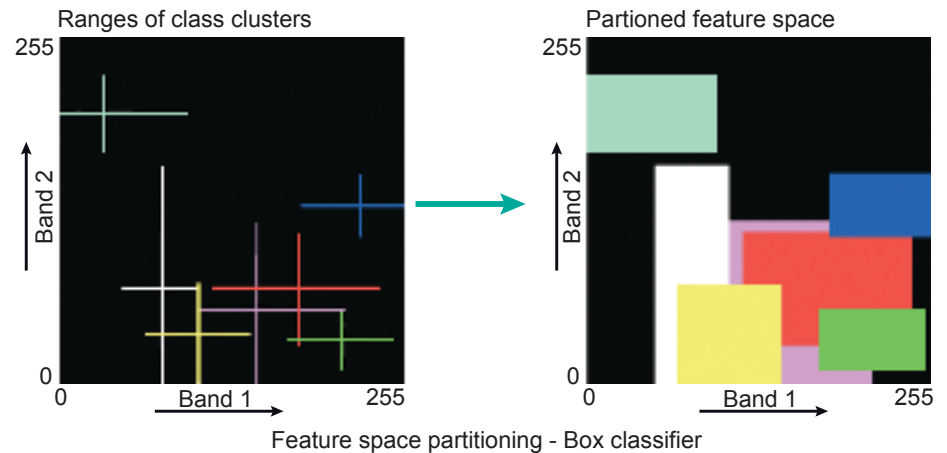
Figure 6.13
Principle of the box classification in a two-dimensional situation.

Classification algorithms

After the training sample sets have been defined, classification of the image can be carried out by applying a classification algorithm. Several classification algorithms exist. The choice of the algorithm depends on the purpose of the classification and the characteristics of the image and training data. The operator needs to decide if a *reject* or *unknown* class is allowed. In the following, three classifier algorithms are described. First the *box classifier* is explained—its simplicity helps in understanding the principle. In practice, the box classifier is hardly ever used, however; *minimum distance to mean* and the *maximum likelihood* classifiers are most frequently used.

Box classifier

The box classifier is the simplest classification method. For this purpose, upper and lower limits are defined for each band and each class. The limits may be based on the minimum and maximum values or on the mean and standard deviation per class. When the lower and the upper limits are used, they define a box-like area in the feature space (Figure 6.13). The number of boxes depends on the number of classes. During classification, every feature vector of an input (two-band) image will be checked to see if it falls in any of the boxes. If so, the cell will get the class label of the box it belongs to. Cells that do not fall inside any of the boxes will be assigned the "unknown class", sometimes also referred to as the "reject class".

The disadvantage of the box classifier is the overlap between classes. In such a case, a cell is arbitrarily assigned the label of the first box it encounters.

Minimum Distance to Mean classifier

The basis for the minimum distance to mean (MDM) classifier is the cluster centres. During classification, the Euclidean distances from a candidate feature vector to all the cluster centres are calculated. The candidate cell is assigned to the class that qualifies as the closest one. Figure 6.14 illustrates how a feature space is partitioned based on the cluster centres. One of the disadvantages of the MDM classifier is that points that are at a large distance from a cluster centre may still be assigned to this centre.

Means of class clusters

Minimum distance to Mean classifier

Partioned feature space
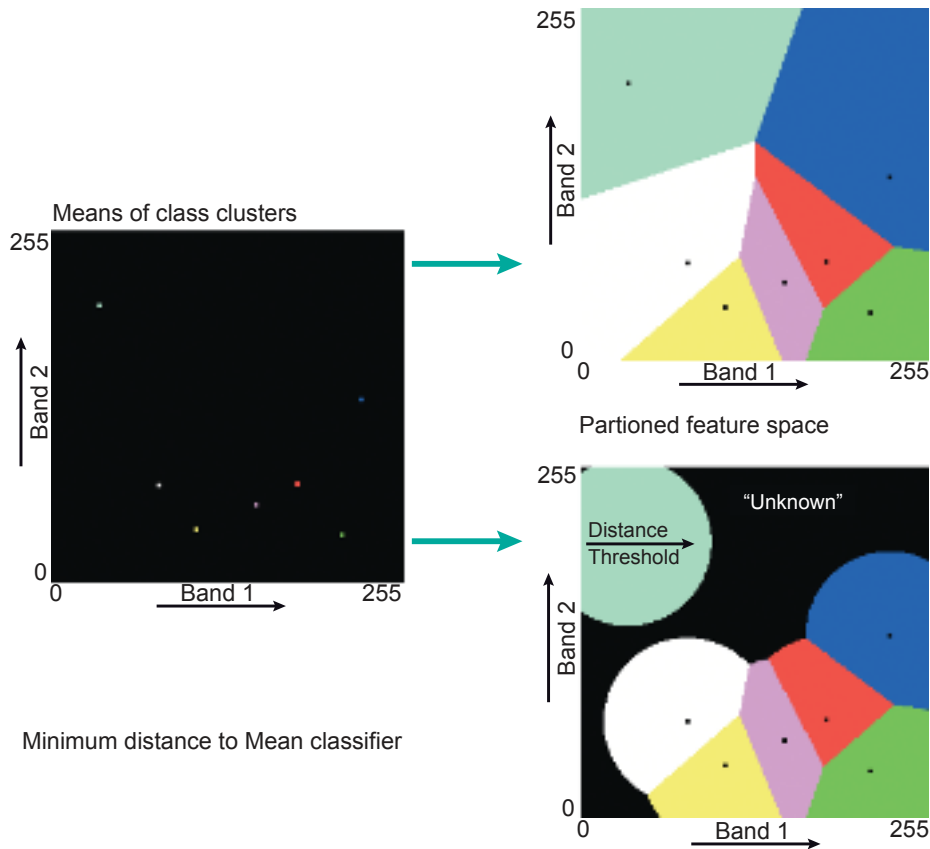
Figure 6.14
Principle of the *minimum distance to mean* classification in a two-dimensional situation. The decision boundaries are shown for a situation without threshold distance (upper right) and one with threshold distance (lower right).

This problem can be overcome by defining a threshold value that limits the search distance. Figure 6.14 illustrates the effect; the threshold distance to the centre is shown as a circle.

A further disadvantage of the MDM classifier is that it does not take the class variability into account: some clusters are small and dense, while others are large and dispersed. Maximum likelihood classification does, however, take class variability into account.

Maximum Likelihood classifier

The Maximum likelihood (ML) classifier considers not only the cluster centres but also the shape, size and orientation of the clusters. This is achieved by calculating a statistical distance based on the mean values and covariance matrix of the clusters. The statistical distance is a probability value: the probability that observation x belongs to specific cluster. A cell is assigned to the class (cluster) for which it has the highest probability. The assumption of most ML classifiers is that the statistics of the clusters follow a *normal* (Gaussian) distribution.

For each cluster, what are known as "equiprobability contours" can be drawn around the centres of the clusters. Maximum likelihood also allows the operator to define a threshold distance by defining a maximum probability value. A small ellipse centred on the mean defines the values with the highest probability of membership of a class. Progressively larger ellipses surrounding the centre represent contours of probability of membership to a class, with the probability decreasing the further away from the

centre. Figure 6.15 shows the decision boundaries for a situation with and without threshold distance.

6.2.4 Validation of the result

Image classification results in a raster file in which the individual raster elements are labelled by class. As image classification is based on samples of the classes, the actual quality of the classification result should be checked. This is usually done by a sampling approach in which a number of raster elements of the output are selected and both the classification result and the "true world class" are compared. Comparison is done by creating an *error matrix* from which different accuracy measures can be calculated. The true world class is preferably derived from field observations. Sometimes, sources for which higher accuracy can be assumed, such as aerial photos, are used as a reference.

sampling schemes

Various *sampling schemes* have been proposed for selecting pixels to test. Choices to be made relate to the design of the sampling strategy, the number of samples required, and the area of the samples. Recommended sampling strategies in the context of land cover data are simple random sampling or stratified random sampling. The number of samples may be related to two factors in accuracy assessment: (1) the number of samples that must be taken in order to reject a data set as being inaccurate; or (2) the number of samples required to determine the true accuracy, within some error bounds, of a data set. Sampling theory is used to determine the number of samples required. The number of samples must be traded-off against the area covered by a sample unit. A sample unit can be a point but it could also be an area of some size; it can be a single raster element but may also include surrounding raster elements. Among other
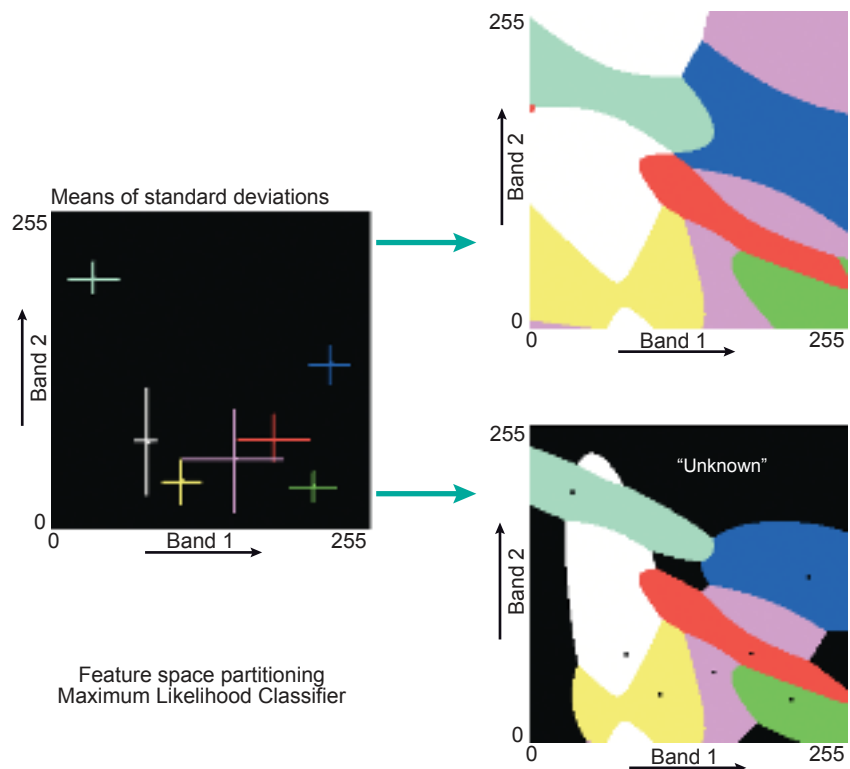
Figure 6.15
Principle of *maximum likelihood* classification. The decision boundaries are shown for a situation without threshold distance (upper right) and one with threshold distance (lower right).

considerations, the "optimal" sample-area size depends on the heterogeneity of the class.

	A	B	C	D	Total	Error of Commission (%)	User Accuracy (%)
a	35	14	11	1	61	43	57
b	4	11	3	0	18	39	61
c	12	9	38	4	63	40	60
d	2	5	12	2	21	90	10
Total	53	39	64	7	163		
Error of Omission	34	72	41	71			
Producer Accuracy	66	28	59	29			

Table 6.2
The error matrix with derived errors and accuracy expressed as percentages. A, B, C and D refer to the reference classes; a, b, c and d refer to the classes in the classification result. Overall accuracy is 53%.

Once sampling for validation has been carried out and the data collected, an error matrix, also sometimes called a *confusion matrix* or an *contingency matrix*, can be established (Table 6.2). In the table, four classes (A, B, C, D) are listed. A total of 163 samples were collected. The table shows that, for example, 53 cases of A were found in the real world ('reference'), while the classification result yielded 61 cases of a; in 35 cases they agree.

The first and most commonly cited measure of mapping accuracy is the *overall accuracy*, or proportion correctly classified (PCC). Overall accuracy is the number of correctly classified pixels (i.e. the sum of the diagonal cells in the error matrix) divided by the total number of pixels checked. In Table 6.2 the overall accuracy is $(35 + 11 + 38 + 2)/163 = 53\%$. The overall accuracy yields one value for the result as a whole.

Most other measures derived from the error matrix are calculated per class. *Error of omission* refers to those sample points that are omitted in the interpretation result. - Consider class A, for which 53 samples were taken. Eighteen out of the 53 samples were interpreted as b, c or d. This results in an error of omission of $18/53 = 34\%$. Error of omission starts from the reference data and therefore relates to the columns in the error matrix. The *error of commission* starts from the interpretation result and refers to the rows in the error matrix. The error of commission refers to incorrectly classified samples. Consider class d: only two of the 21 samples (10%) are correctly labelled. Errors of commission and omission are also referred to as Type I and Type II errors, respectively.

Omission error is the corollary of producer accuracy, while user accuracy is the corollary of commission error. The user accuracy is the probability that a certain reference class has indeed actually been labelled as that class. Similarly, producer accuracy is the probability that a sampled point on the map is indeed that particular class.

Another widely used measure of map accuracy derived from the error matrix is the kappa or κ coefficient. Let x_{ij} denote the element of the error matrix in row i and column j, r denote number of classes and N total sum of all elements of the error matrix. Then kappa coefficient is computed as

$$\kappa = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i+} x_{+i}}{N^2 - \sum_{i=1}^{r} x_{i+} x_{+i}} \tag{6.1}$$

where $x_{i+} = \sum_{j=1}^{r} x_{ij}$ and $x_{+i} = \sum_{j=1}^{r} x_{ji}$ are the sums of all elements in row i and column i, respectively.

Kappa coefficient takes into account the fact that even assigning labels at random will result in a certain degree of accuracy. Kappa statistics, based on kappa coefficient, can

be applied to test if two data sets, e.g. classification results, have different levels of accuracy. This type of testing is used to evaluate different RS data or methods for the generation of spatial data.

6.2.5 Pixel-based and object-oriented classification

Pixel-based image classification is a powerful technique to derive *thematic classes* from multi-band images. However, it has certain limitations that users should be aware of. The most important constraints of pixel-based image classification are that it results in (i) spectral classes, and that (ii) each pixel is assigned to one class only.

Spectral classes are those that are directly linked to the spectral bands used in the classification. In turn, these are linked to surface characteristics. In that respect, one can say that spectral classes correspond to land cover classes. In the classification process a *spectral class* may be represented by several *training classes*. Among other things, this is due to the variability within a spectral class. Consider, for example, a class such as "grass": there are different types of grass, each of which has different spectral characteristics. Furthermore, the same type of grass may have different spectral characteristics when considered over larger areas, owing to, for example, different soils and climatic conditions.

spectral classes

A related issue is that sometimes one is interested in land use classes rather than land cover classes. Sometimes, a land use class may comprise several land cover classes. Table 6.3 gives some examples of links between spectral land cover and land use classes. Note that between two columns there can be 1-to-1, 1-to-n, and n-to-1 relationships. The 1-to-n relationships are a serious problem and can only be solved by adding data and/or knowledge to the classification procedure. The data added can be other remote sensing images (other bands, other moments) or existing geospatial data, such as topographic maps, historical land inventories, road maps, and so on. Usually this is done in combination with adding expert knowledge to the process. An example would be using historical land cover data and defining the probability of certain land cover changes. Another example would be to use elevation, slope and aspect information. This will prove especially useful in mountainous regions, where elevation differences play an important role in variations of surface-cover types.

Table 6.3
Spectral classes distinguished during classification can be aggregated to land cover classes. 1-to-n and n-to-1 relationships can exist between land cover and land use classes.

Spectral class	Land cover class	Land use class
water	water	shrimp cultivation
grass1	grass	nature reserve
grass2	grass	nature reserve
grass3	grass	nature reserve
bare soil	bare soil	nature reserve
trees1	forest	nature reserve
trees2	forest	production forest
trees3	forest	city park

The other main problem and limitation of pixel-based image classification is that each pixel is only assigned to one class. This is not a problem when dealing with relatively small ground resolution cells. However, when dealing with relatively large GRCs, more land cover classes are likely to occur within a cell. As a result, the value of the pixel is an average of the reflectance of the land cover present within the GRC. In a standard classification, these contributions cannot be traced back and the pixel will be assigned to one of either classes or perhaps even to another class. This phenomenon is usually referred to as the *mixed pixel*, or *mixel* (Figure 6.16). This problem of mixed pixels is inherent to image classification: assigning a pixel to one thematic class. The

mixed pixels
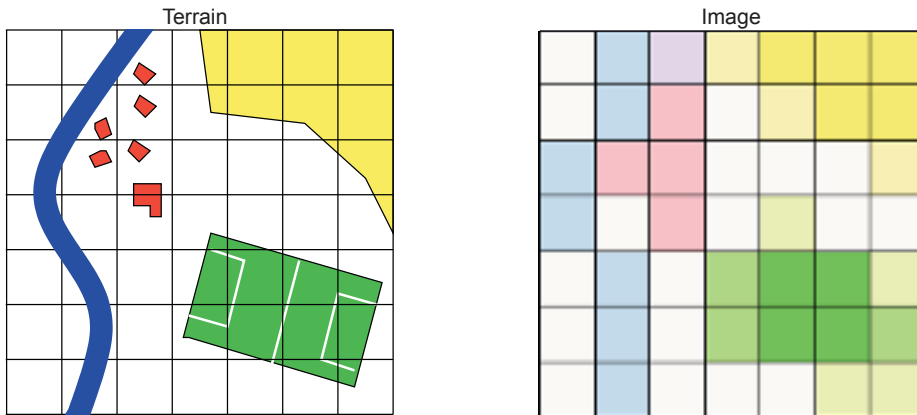
Terrain   Image

solution to this is to use a different approach, for example, by assigning the pixel to more than one class. This brief introduction into the problem of mixed pixels also highlights the importance of using data with the appropriate spatial resolution.

We have seen that the choice of classification approach depends on the data available, but also on the knowledge we have about the area under investigation. Without knowledge of the land cover classes present, unsupervised classification can only give an overview of the variety of classes in an image. If knowledge is available, from field work or other sources, supervised classification may be superior. However, both methods only make use of spectral information, which becomes increasingly problematic for higher spatial resolutions. For example, a building that is made up of different materials leads to pixels with highly variable spectral characteristics, and thus a situation for which training of pixels is of little help. Similarly, a field may contain healthy vegetation pixels as well as some of bare soil.

We are also increasingly interested in land use. However, to distinguish, for example, urban from rural woodland, or a swimming pool from a natural pond, an approach similar to visual interpretation (as described in Section 6.1) is needed. Object-oriented analysis (OOA), also called segmentation-based analysis, allows us to do that. Instead of trying to classify every pixel separately, and only based on spectral information, OOA breaks down an image into spectrally homogenous segments that correspond to fields, tree stands, buildings, etc. It is also possible to use auxiliary GIS layers, for example building footprints, to guide this segmentation. Similarly to the cognitive approach of visual image interpretation—where we consider each element in terms of its spectral appearance but also in terms of its shape and texture, and within its environment—in OOA we can then specify contextual relationships and more complex segment characteristics to classify the objects extracted in the segmentation process. For example, we can use object texture to distinguish two spectrally similar forest types, or distinguish a swimming pool from a pond, by considering its shape and perhaps the surrounding concrete instead of soil and vegetation. OOA is particularly suitable for images of high spatial resolution, but also for data obtained by ALS or microwave radar. It requires that we have substantial knowledge on what distinguishes a given land cover or land use type, as well as auxiliary data such as elevation, soil type or vector layers.

object-oriented analysis

Chapter 7

Models and modelling

Rolf de By
Menno-Jan Kraak
Otto Huisman

Introduction

The working environment of the geoscientist should offer functions to deal with the diversity of the data at any moment is what can be called geodata processing. It should encourage connections between the different phases of problem solving and facilitate the linking of concepts, data and models. Geodata processing comprises the steps of exploration, synthesis, analysis, evaluation and presentation, although not necessarily in that sequential order (Figure 7.1). In theory, with a particular geo-problem in mind, one starts with the exploration of available data, which leads to a research hypothesis or a conceptual idea.

Synthesis, the next step, leads to theories or models that will be subject to analytical processes. The results of these will be evaluated and followed up with the presentation of the final result, indicating the confidence one has in the result, so that one might may appropriate decisions. In practice, geodata processing might not involve all of these steps or even consist of several iterations.

Modelling is commonly carried out to solve real-world problems. A modelling activity must therefore begin with a clearly-stated purpose. Sometimes, is more useful to discuss the modelling process itself rather than the precise meaning of the model. A model can be defined as a simplified representation of some aspects of a real system. By creating a model, we move from the real world into an abstract world of concepts, mathematical constructs, spatial queries and techniques for delivering solutions. We re-enter the real world with the solution in hand, at which stage that solution is translated into a useful answer to the real-world problem. Models—as simplified representations—come in many different "flavours".

7.1 What is a model?

The word "model" can be read as a verb, which means "to describe" or "to represent", but it can also be used as a noun. In Van Dale's Dutch dictionary (Van Dale's Groot Woordenboek der Nederlandse Taal) a "model" is defined as "a schematization of
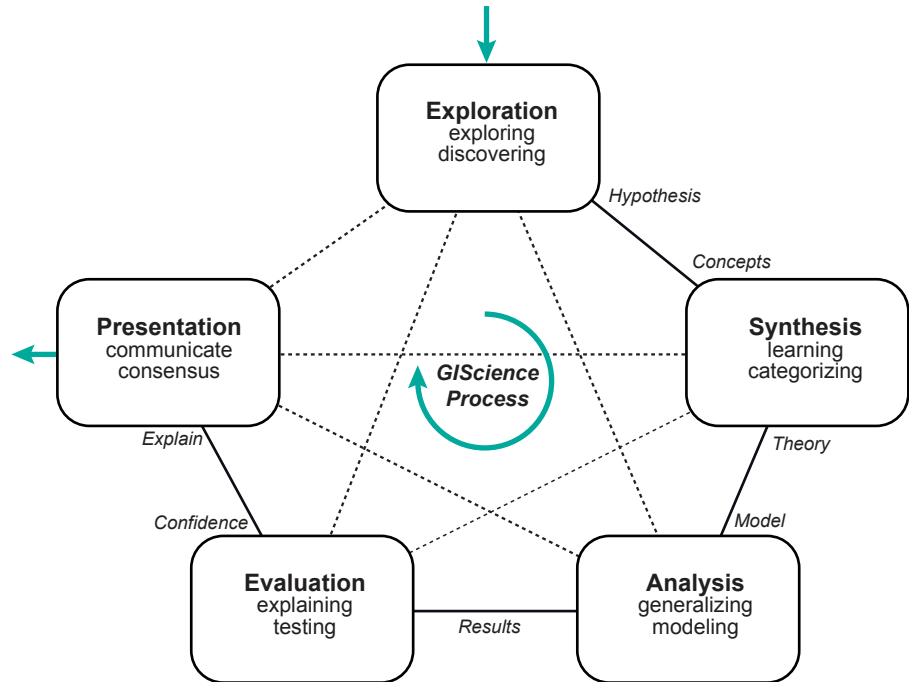
Figure 7.1
Geodata processing comprises the steps of exploration, synthesis, analysis, evaluation, and presentation. Between each of these steps, certain actions—via a combination of visual an algorithmic approaches—are executed.

model

reality with operational potentials". This points to the intimate relationship between the type and degree of schematization and the objectives for the call on the operational potentials. Meijer (1984) stresses this call on potentials as follows: "A model is an object or concept that is used to represent something else. It is reality scaled down and converted to a form we can comprehend."

Here we will use the following definition of a model: "A model is a manageable, comprehensible and schematic representation of a piece of reality". This definition can be explained as follows:

- "reality": as scientists we usually wish to model reality and not a hypothetical system. Reality is, however, vast and infinitely complex. Describing (modelling) reality in all its facets and complexity is therefore impossible.

- "a piece of reality": we focus on a limited domain in time and space, incorporating only specific aspects, focusing on perhaps only one or just a few objectives. We represent a number of phenomena that we can observe in reality, usually to enable study, administration, computation and/or simulation.

- "schematic representation": reality is only considered from a specific point of view. This implies imperfections and limited accuracy (e.g. projection of the Earth onto a map).

- "representation": an infinite number of projections, thus different representations of the same piece of reality are possible, resulting in an equal number of modelling possibilities.

- "a comprehensible representation": a model serves a specific goal. A model is used as an aid to diagnose, analyse or predict the behaviour of the modelled system. The user is central in this. The user needs to be able to understand the

functioning of the model itself, as well as how the model and the real world differ.

- "a manageable representation": for the same reason as the previous item, the model needs to be manageable. It should give users the results they need to be able to answer their questions, to find solutions to their problems. Users need to be able to "play" with the model (change variables, experiment, evaluate, etc.).

So far, we have not discussed types of models, which may be mathematical or statistical models, conceptual models, spatial models, scale models, to mention a few. Before we look closer at model types, we first need to discuss the various functions of models.

7.2 Function and use of models

Models can be used to serve various objectives or functions, for example to better understand a spatial phenomenon, to represent a phenomenon, to measure or calculate a distance, or to set up plans for a disaster response unit within a city.

7.2.1 Insight versus prediction

Models may be used to gain a better understanding of the system that they represent, i.e. the structure of real-world systems and the way they work. Different elements of the model and their interrelations may be manipulated by changing variables, attributes, equations and parameters. As such, it is possible to experiment using a model instead of with the real-world system (even if that would at all be possible, simulating flooding under real-world conditions is usually not possible or feasible). These kind of models are used abundantly in science.

Models can also be used to predict the future behaviour of a system, with or without interventions. It is important is to ensure that the model represents future behaviour accurately enough (usually based on knowledge about current behaviour or historical trends). Simple extrapolations may not always be desirable, as results obtained from the past are no guarantee for future behaviour!

Scenarios based on simulations may be used to sketch possible future developments. A possible future is constructed based on a well-described initial situation. Over time, several changes occur or are introduced, thereby changing the predictions. Scenarios are particularly useful when used in comparison with other scenarios (starting from the same initial situation). Scenario models and model predictions are typically used in policy studies.

7.2.2 Measuring versus calculating

In many applications it is not sufficient to only know what is happening (qualitative). Usually we also need to know to what extent something is happening (quantitative). Such quantification is common in, for example, engineering, finance and economics, environmental sciences and physical geography. Quantitative measurements can be obtained by direct measurement (from prototypes, scale models or analogue models) or by calculation. In the latter case, scientists need quantitative information to answer questions such as "What traffic volumes are to be expected on this new bridge?" or "How large is this deforested area?" In the first question, the object or system has not yet been realized if the bridge is not already built and operating, so only a predictive model can help. If such a model is a mathematical model, then the different variables (quantities) that are considered to be important and their interrelationships

are expressed in equations (formulas), after which mathematical operations can be executed (to predict, to control, to optimize). In the second question, the deforestation has already taken place and direct measurement is possible (e.g. through a NDVI).

7.3   Modelling

It is not possible for we human beings to know or comprehend reality completely. Nevertheless we create images of reality that differ both between individuals and over time. These images often consist of a simplified reproduction of the processes in or around us, the relations between these processes, and so on. In creating an image, it is impossible to involve every aspect of a process simultaneously. Fortunately, we do not need to do so: it is sufficient to consider only a few aspects at a time. Moreover, we usually create an image as a schematic reproduction. Which aspects we involve and to what extent we use schematic reproductions depends on, for example, our activities, knowledge, interests, disposition and the required precision.

Models are constructed to answer questions concerning parts of reality. This restricts their use to only certain aspects of the modelled phenomena. Moreover, models are limited in time and space. A model will, therefore, never be universal. A third important characteristic of models is that they are—even within the limits of the chosen aspect(s)—only an approximation; anything not primarily important will be left out. Remember, models are a schematic representation of a piece of reality.

These characteristics make clear the limited validity of models. If many factors are involved, it is especially difficult to indicate the validity—within which errors will be reduced to an acceptable amount—in advance. The implementation of models therefore involves uncertainty, so validation is needed, i.e. testing the model to see if it has a sufficient sense of reality.

uncertainty

The schematic representation of a model implies that for a certain question, there is no one correct model. On the contrary, a multitude of models could be relevant for the question concerned. Depending on the required precision and the time and means available, we may want to use a simple model or a complicated one, representing more aspects or more detail. The question "which model do we need to solve which problem?" is therefore inappropriate since many models may do the job.

In the process of developing a model, several concepts, terms and variables that may not be literally part of reality are introduced. Those concepts and variables represent certain idealized aspects of this reality and are called modelled variables.

Applying a model following recognized (mathematical) rules generates model outcomes. For these outcomes, the same rules apply as for the modelled variables: they are valid *within* the model and should not be seen as characteristics of the modelled reality. This is similar to the rules and outcomes of a board game: they are limited to the game and do not apply to the real world.

Because each model is based on a schematic representation of reality, or some part of it, the outcomes may be unrealistic when compared to the reality the model wants to represent. This requires a critical evaluation of models and their outcomes, even if a model has been previously validated; it is not as if its implications automatically involve the same situation. Those who implement a model therefore need sufficient understanding and knowledge of the processes being modelled in order to be able to judge the model. Unfortunately, this is not always the case. The easy accessibility of computer programs has given modelling capabilities to people who may not be aware of the limits of their knowledge or the validity of the model outcomes.

7.4 General characteristics of models

A few general characteristics apply to all models. When building a model the following questions could be useful for exploring the model's general specifications/attributes:

- what is the problem to be modelled? (e.g. inaccessibility of a city centre);

- what are the important phenomena? (e.g. road congestion, urban land use);

- what is the spatial domain? (e.g. a built-up area of a city);

- what is the temporal domain? (e.g. morning peak-hour);

- what is the desired accuracy? Depending on the type of use, more- or less-detailed modelling (e.g in terms of spatial resolution) may be needed.

For operational use of the model, other questions are also relevant, such as:

- who is going to use the model (the problem owner?); and

- what resources are available for building the model (time, money).

Randers [96] distinguished nine general characteristics of models, where the objective of the study determines the importance of each of the characteristics. These general characteristics are:

- insight-generating capacity: the model increases insight into the system and its image;

- descriptive realism: system and model structure are similar;

- model reproduction ability: model reproduces typical system behaviour;

- transparency: the model is useable and understandable, also by non-experts;

- relevance: according to experts, the model provides meaningful solutions to the problem;

- ease of enrichment: the model can easily be adapted;

- fertility: the model stimulates generation of new ideas, new experiments, new policy;

- formal correspondence with data: the model reproduces known observations accurately;

- point predictive ability: the model makes good predictions.

Most models, particularly those used in GI Science, try to increase knowledge of and insight into the observed behaviour of systems or to estimate the effect of changes (choice options) on their behaviour.

7.5 How to build a model

The modelling process consists of three distinct phases: model development, model operationalization, and model application, in that order.
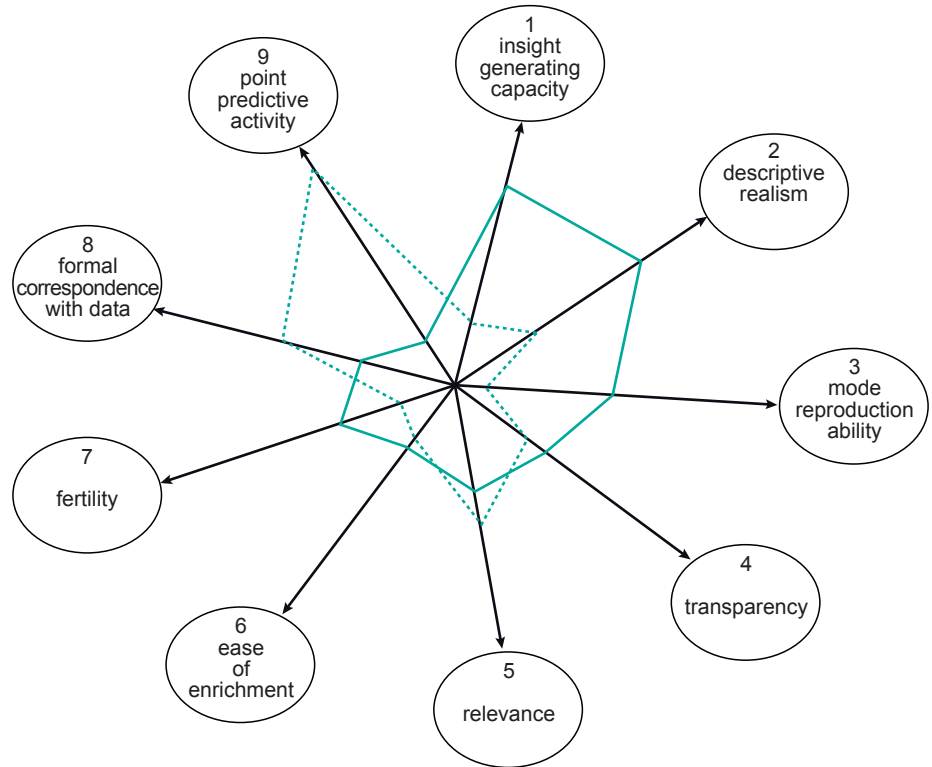
Figure 7.2
A spider diagram relating
model characteristics
according to Randers [96]
and the study objective.
Broken lines indicate
predictive types of models;
solid lines indicate descriptive
types of models.

Model development   The following 10 steps (see Figure 7.3) may lead to the proper
(mathematical) development of a model:

- problem analysis;

- conceptual modelling;

- model formulation;

- conceptual validation;

- model implementation;

- verification;

- calibration;

- validation;

- analysis; and

- model use, communication and evaluation.

Operationalizing   In the second phase, modelling concept and modelling are im-
plemented, perhaps using commercially-available software. Functional and technical
design, software validation, user-interface development, etc., are addressed. The com-
pilation of a manual with a detailed description of the underlying model, including
metadata, belongs in the operationalization process.
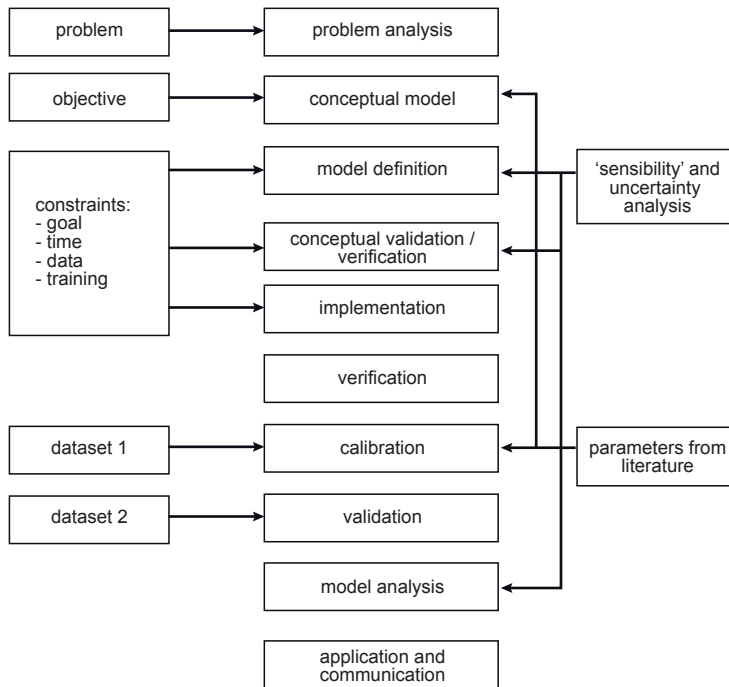
Figure 7.3
A summary of the modelling
process.

Application   In the third phase, the model is used for calculations, simulations or making predictions within the context of the problem that had to be solved in the first place.

In total, the modelling process therefore entails the following activities: problem analysis (including area demarcation; schematic representation of processes; choice of modelling type and strategy; data collection, model development; and model calibration and verification. Sensitivity and uncertainty analysis, model presentation and visualization, and, finally, model communication may also be part of this process.

7.6   Modelling in GISs

In the GIS environment, the most familiar model is a map. A map is a miniature representation of a part of the real world. Databases are another important class of models. A database contains (spatial) data and also provides various functions for operating on that data. The collection of stored data represents a real-world phenomena, so it also is a model. Digital models (such as a database or GISs) have advantages over paper models (e.g. analogue maps), being more flexible and therefore more adaptable for the purpose at hand. Digital models may also allow animations and simulations.

"Data modelling" is the common name for the design effort of structuring a database. It involves the identification of the kinds of data that the database will store and the relationships between them. Most maps and databases can be considered static models, representing a single state of affairs at a point in time. Dynamic models or process models, on the other hand, address developments or changes in the real world: changes in the past, in the present or at sometime in the future. Dynamic models are inherently more complicated than static models and usually require much more computation. Simulation models are an important class of dynamic models that allow the

simulation of real-world processes.

GISs and application models   We have already mentioned that real-world processes are often highly complex. Models are simplified abstractions of reality representing or describing its most important elements and their interactions. Modelling and GISs are connected, since a GIS is itself a tool for modelling a part of "the real world". The solution to a problem usually depends on a number of parameters. Since these parameters are often interrelated, their interaction is made more precise in an application model. Such a model describes as faithfully as possible how the relevant phenomena (in our case, often geographic) behave, doing so in terms of the parameters. Application models varying in nature. GIS applications for famine-relief programmes, for instance, are different from earthquake-risk assessment applications, though both use GISs to derive a solution. Many kinds of application models exist and they can usually be classified in various ways. Here we identify five ways of classifying the characteristics of GIS-based application models (compare with the model of Randers in Figure 7.2):

- the purpose of the model;

- the methodology underlying the model;

- the scale at which the model works;

- its dimensionality, i.e. whether the model includes spatial, temporal or spatial and temporal dimensions;

- its implementation logic, i.e. the extent to which the model uses existing knowledge about the implementation context.

These classifications express different ways for classifying model characteristics. Any application model has characteristics that can be described according to the above categories. Each of these is briefly discussed here.

Purpose of the model   The purpose of a model refers to whether the model is descriptive, prescriptive or predictive in nature. Descriptive models attempt to answer the "what is" question. Prescriptive models answer the "what should be" question by determining the best solution from a given set of conditions. Models for planning and site selection are prescriptive in that they quantify environmental, economic and social factors to determine "best" or optimal locations. Predictive models focus upon the "what is likely to be" questions and predict outcomes based upon a set of input conditions. Examples of predictive models include forecasting models, such as those attempting to predict landslides or sea level rise.

Methodology underlying the model   The underlying methodology of a model refers to its operational components. Stochastic models use probabilistic functions to represent random or semi-random behaviour of phenomena. By contrast, deterministic models are based upon a well-defined cause–effect relationship. Examples of deterministic models include hydrological flow and pollution models, where the "effect" can often be described by numerical methods and differential equations. Rule-based models address processes by using local (spatial) rules. Cellular Automata (CA), for example, are often used for systems that are generally not well understood, but for which local processes are well known. For example, to model the direction of spread

of a fire over several time-steps, the characteristics of neighbouring cells (such as wind direction and vegetation type) in a raster-based CA model might be used. Agent-based models (ABMs) model movement and development of multiple interacting agents (which might represent individuals), often using sets of decision rules about what the agent can and cannot do. Complex agent-based models have been developed to understand aspects of travel behaviour and crowd interactions, incorporating also stochastic components.

Scale   Scale refers to whether a model component is individual or aggregate in nature. It refers to the "level" at which the model operates. Individual model components are based on individual entities, such as in agent-based models, whereas aggregate models deal with "grouped" data, such as population census and socio-economic data. Models may operate on data at the level of a city block (for example, using population census data for particular social groups), at a regional level or even a global one.

Dimensionality   Dimensionality refers to the static or dynamic character of a model and to its spatial or non-spatial nature. Static vs. dynamic modelling has been discussed in Subsection 7.6. Models operating in a geographically defined space are explicitly spatial, whereas models without spatial reference are aspatial.

Implementation logic   Implementation logic refers to how the model uses existing theory or knowledge to create new knowledge. Deductive approaches use knowledge of the overall situation in order to predict outcome conditions. This includes models that have a formalized set of criteria, often with known weightings for the inputs, and for which existing algorithms are used to derive outcomes. Inductive approaches, on the other hand, are less straightforward, in that they try to generalize (often based upon samples of a specific data set) in order to derive more general models. While an inductive approach is useful if we do not know the general conditions or rules that apply to a specific domain, it is typically a trial and error approach that requires empirical testing to determine the parameters of each input variable. Most GISs have a limited range of tools for modelling. For complex models, or functions that are not natively supported in a GIS, external software environments are frequently used. In some cases, GISs and models can be fully integrated (known as embedded coupling) or linked through their data and interface (known as tight coupling). If neither is possible, the external model might run independently of a GIS; the model output should be exported into the GIS for further analysis and visualization. This is known as loose coupling.

Chapter 8

Spatial data modelling, collection and management

Rolf de By
Otto Huisman
Richard Knippers
Menno-Jan Kraak
Alfred Stein

8.1 Geographic Information and spatial data types

8.1.1 Geographic phenomena

Defining geographic phenomena

A GIS operates under the assumption that the spatial phenomena involved occur in a two- or three-dimensional Euclidean space. Euclidean space can be informally defined as a model of space in which locations are represented by coordinates—(x, y) in 2D and (x, y, z) in 3D space—and distance and direction can defined with geometric formulas. In 2D, this is known as the Euclidean plane. To represent relevant aspects of real-world phenomena inside a GIS, we first need to define what it is we are referring to. We might define a geographic phenomenon as a manifestation of an entity or process of interest that:

Euclidean space

geographic phenomena

- can be named or described;

- can be georeferenced; and

- can be assigned a time (interval) at which it is/was present.

Relevance of phenomena for the use of a GIS depends entirely on the objectives of the study at hand. For instance, in water management, relevant objects can be river

basins, agro-ecological units, measurements of actual evapotranspiration, meteorological data, groundwater levels, irrigation levels, water budgets and measurements of total water use. All of these can be named or described, georeferenced and provided with a time interval at which each exists. In multipurpose cadastral administration, the objects of study are different: houses, land parcels, streets of various types, land use forms, sewage canals and other forms of urban infrastructure may all play a role. Again, these can be named or described, georeferenced and assigned a time interval of existence.

objectives of the application

Not all relevant information about phenomena has the form of a triplet (description, georeference, time interval). If the georeference is missing, then the object is not positioned in space: an example of this would be a legal document in a cadastral system. It is obviously somewhere, but its position in space is not considered relevant. If the time interval is missing, we might have a phenomenon of interest that exists permanently, i.e. the time interval is infinite. If the description is missing, then we have something that exists in space and time, yet cannot be described. Obviously this last issue limits the usefulness of the information.

Types of geographic phenomena

The definition of geographic phenomena attempted above is necessarily abstract and is, therefore, perhaps somewhat difficult to grasp. The main reason is that geographic phenomena come in different "flavours". Before categorizing such "flavours", there are two further observations to be made.

First, to represent a phenomenon in a GIS requires us to state what it is and where it is. We must provide a description—or at least a name—on the one hand, and a georeference on the other hand. We will ignore temporal issues for the moment and come back to these in Subsection 8.1.4, the reason being that current GISs do not provide much automatic support for time-dependent data. This topic must, therefore, be considered as an example of advanced GIS use. Second, some phenomena are manifest throughout a study area, while others only occur in specific localities. The first type of phenomena we call geographic fields; the second type we call objects.

Geographic fields

A geographic field is a geographic phenomenon that has a value "everywhere" in the study area. We can therefore think of a field as a mathematical function f that associates a specific value with any position in the study area. Hence if (x, y) is a position in the study area, then $f(x, y)$ expresses the value of f at location (x, y). Fields can be discrete or continuous. In a continuous field, the underlying function is assumed to be "mathematically smooth", meaning that the field values along any path through the study area do not change abruptly, but only gradually. Good examples of continuous fields are air temperature, barometric pressure, soil salinity and elevation. A continuous field can even be differentiable, meaning that we can determine a measure of change in the field value per unit of distance anywhere and in any direction. For example, if the field is elevation, this measure would be slope, i.e. the change of elevation per metre distance; if the field is soil salinity, it would be salinity gradient, i.e. the change of salinity per metre distance.

discrete and continuous fields

Figure 8.1 illustrates the variation in elevation of a study area in Falset, Spain. A colour scheme has been chosen to depict that variation. This is a typical example of a continuous field. Discrete fields divide the study space in mutually exclusive, bounded parts, with all locations in one part having the same field value. Typical examples are land classifications, for instance, using either geological classes, soil type, land use type, crop type or natural vegetation type. An example of a discrete field—in this case iden-

tifying geological units in the Falset study area—is provided in Figure 8.2. Observe that locations on the boundary between two parts can be assigned the field value of the "left" or "right" part of that boundary.

Elevation in the Falset study area, Tarragona province, Spain. The area is approximately 25x20 km.
The illustration has been easthetically improved by a technique known as "hillshading". In this case, it is as if the sun shines from the north-west, giving a shadow effect towards the south-east. Thus, colour alone is not a good indicator of elevation; observe that elevation is a coninuous function over the space.

Figure 8.1
An example of a continuous field, namely the *elevation* in a study area in Falset, Spain.

Discrete fields are intermediate between continuous fields and geographic objects: discrete fields and objects both use "bounded" features. A discrete field, however, assigns a value to every location in the study area, which is not typically the case for geographic objects. These two types of fields differ in the type of cell values. A discrete field such as land use type will store cell values of the type "integer" and is therefore also called an integer raster. Discrete fields can be easily converted to polygons since it is relatively easy to draw a boundary line around a group of cells with the same value. A continuous raster is also called a "floating point" raster.

A field-based model consists of a finite collection of geographic fields: we may be interested in, for example, elevation, barometric pressure, mean annual rainfall and maximum daily evapotranspiration, and would therefore use four different fields to model the relevant phenomena within our study area.

Data types and values   Different kinds of data values may represent spatial "phenomena". Some of these kinds of data limit the types of analyses that can be done on the data:

1. Nominal data values provide a name or identifier to discriminate between different values. Specifically, we cannot do true computations with these values. An example is the names of geological units. This kind of data value is called categorical when the values assigned are sorted according to a set of non-overlapping categories. For example, we might identify the soil type of a given area as belonging to a certain (pre-defined) category.

2. Ordinal data values are numerical data that can be put in a natural sequence but that do not allow any other type of computation. Household income, for instance, could be classified as being either "low", "average" or "high". Clearly this is their natural sequence, but this is all we can say—we can not say that a high income is twice as high as an average income.

3. Interval data values are numerical data that allow simple forms of computation like addition and subtraction. However, interval data have no arithmetic zero value, and do not support multiplication or division. For instance, a temperature of 20 °C is not twice as hot as 10 °C.

4. Ratio data values are numerical data that allow most, if not all, forms of arithmetic computation. Ratio data have a natural zero value and multiplication and division of values are possible operators (distances measured in metres are an example of ratio data). Continuous fields can be expected to have ratio data values, hence we can interpolate them.

☐ Miocene and Quaternay (lower left)	☐ Lias
☐ Oligocene (left)	☐ Keuper and Muschelkalk
☐ Cretaceous (right)	☐ Bundsandstein
☐ Eocene	☐ Intrusive and sedimentary areas

Observe that - typical for fields - with any location only a single geological unit is associated. As this is a *discrete* field, value changes are discontinuous, and therefore locations on the boundary between two units are *not* associated with a particular value (i.e. with a geological unit).

Figure 8.2
A discrete field indicating geological units as used in a foundation-engineering study for constructing buildings. Falset study area as in Figure 8.1.

We usually refer to nominal or categorical data values as "qualitative" data because we are limited in terms of the computations we can do on this type of data. Interval and ratio data are known as *quantitative* data as they refer to quantities. Ordinal data, however, do not fit either of these data types. Often, ordinal data refer to a ranking scheme or some kind of hierarchical phenomena. Road networks, for example, are made up of motorways, main roads and residential streets. We might expect roads

qualitative and quantitative data

classified as motorways to have more lanes and carry more traffic than a residential street.

Geographic objects

When a geographic phenomenon is not present everywhere in the study area, but somehow "sparsely" populates it, we look at it as a collection of geographic objects. Such objects are usually easily distinguished and named, and their position in space is determined by a combination of one or more of the following parameters: location (where is it?), shape (what form does it have?), size (how big is it?) and orientation (in which direction is it facing?). How we want to use the information determines which of these four parameters is required to represent the object. For instance, for geographic objects such as petrol stations all that matters in an in-car navigation system is *where* they are. Thus, in this particular context, location alone is enough, and shape, size and orientation are irrelevant. For roads, however, some notion of location (where does the road begin and end?), shape (how many lanes does it have?), size (how far can one travel on it?) and orientation (in which direction can one travel on it?) seem to be relevant components of information in the same system.

geographic objects

Shape is an important component because one of its factors is dimension. This relates to whether an object is perceived as a point feature or a linear, area or volume feature. In the above example, petrol stations are apparently zero-dimensional, i.e. they are perceived as points in space; roads are one-dimensional, as they are considered to be lines. In another use of road information—for instance, in multi-purpose cadastral systems, in which the precise location of sewers and manhole covers matters—roads might be considered as two-dimensional entities, i.e. as areas.

Figure 8.3 illustrates geological faults in the Falset study area, a typical example of a geographic phenomenon that is made up of objects. Each of the faults has a location, and the fault's shape is represented as a one-dimensional object. The size, which is length in the case of one-dimensional objects, is also indicated. Orientation does not play a role here.

Usually, we do not study geographic objects in isolation, but instead look at collections of objects, which we consider as a unit. These collections may have specific geographic characteristics. Most of the more interesting ones obey specific natural laws. The most common (and obvious) of these characteristics is that different objects do not occupy the same location. This, for instance, holds for the collection of petrol stations in an in-car navigation system, the collection of roads in that system, and the collection of land parcels in a cadastral system. We will see in Section 8.1.2 that this natural law of "mutual non-overlap" has been a guiding principle in the design of computer representations of geographic phenomena.

Collections of geographic objects can also be interesting phenomena at a higher level of aggregation: forest plots form forests, groups of parcels form suburbs, streams, brooks and rivers form a river drainage system, roads form a road network, and SST buoys form an SST sensor network. It is sometimes useful to view geographic phenomena at this more aggregated level and look at characteristics such as coverage, connectedness and capacity. For example:

geographic scale

- Which part of a particular road network is within 5 km of a petrol station (a coverage question)?

- What is the shortest route between two cities via the road network (a connectedness question)?

- How many cars can optimally travel from one city to another in an hour (a

capacity question)?

In this context, multi-scale approaches are sometimes used. Such approaches deal with maintaining, and operating on, multiple representations of the same geographic phenomenon, e.g. a point representation in some cases, and an area representation in others. To support these approaches, the database must store representations of geographic phenomena (spatial features) in a scaleless and seamless manner. Scaleless means that all coordinates are world coordinates, i.e. are given in units that are used to reference features in the real world (using a spatial reference system). From such values, calculations can be easily performed and an appropriate scale can be chosen for visualization. Other spatial relationships between the members of a collection of geographic objects may exist and can be relevant in GIS usage. Many of these fall under the category of topological relationships, discussed in Subsection 8.1.3.

Figure 8.3
A number of geological faults in the Falset (Spain) study area; see Figure 8.1. Faults are indicated as blue lines; the study area, with main geological eras, is indicated by the grey background only as a reference.

Boundaries

Where shape or size of areas matter, the notion of a boundary comes into play. This concerns geographic objects but also the constituents of a discrete geographic field, as clearly demonstrated in Figure 8.2. Location, shape and size are fully determined if we know an area's boundary, and thus the boundary is a good candidate for its representation. This especially applies to areas with naturally crisp boundaries. A crisp boundary is one that can be determined at an almost arbitrary level of precision, dependent only on the data-acquisition technique applied. Fuzzy boundaries contrast with crisp boundaries in that a fuzzy boundary is not a precise line, but is rather, itself an area of transition.

As a rule of thumb, crisp boundaries are more common in man-made phenomena, whereas fuzzy boundaries are more common in natural phenomena. In recent years, various research efforts have addressed the issue of explicit treatment of fuzzy bound-

aries, but there is still only limited support in existing GIS software. Typically, the areas identified in a geological classification, like that of Figure 8.2, are vaguely bounded in reality, but applications of this geological information probably do not require high positional accuracy of the boundaries involved. Therefore, an assumption that they are actually crisp boundaries will have little influence on the usefulness of the data

8.1.2 Computer representations of geographic information

Up to this point, we have not considered how geoinformation, such as fields and objects, is represented in a computer. Now that we have discussed the main characteristics of geographic phenomena (above), let us now examine representation in more detail.

Various geographic phenomena have the characteristics of continuous functions in space. Elevation, for instance, can be measured at many locations and each location may give a different value. To represent such a phenomenon in computer memories, we could either:

- try to store as many (location, elevation) observation pairs as possible, or

- try to find a symbolic representation of the elevation field function as a formula in terms of x and y—like $(3.0678x^2 + 20.08x - 7.34y)$ or some such—that can be evaluated to give us the elevation at any given $(x,\ y)$ location.

Both approaches have their drawbacks. A drawback of the first approach is that it is impossible to store all elevation values for all locations since there are infinitely many pairs. A drawback of the second approach is that it is impossible to know the shape of this function, and it would be extremely difficult to derive such a function. In GISs, usually a combination of both approaches is taken. We store a finite, but intelligently chosen set of (sample) locations together with their elevations. This gives us the elevations at the locations stored. An interpolation function allows us to infer an acceptable elevation value for locations that are not stored.

Interpolation is made possible by a principle called spatial autocorrelation. This is a fundamental principle based on Tobler's first law of geography, which states that locations that are closer together are more likely to have similar values than locations that are farther apart. An example is sea-surface temperature, for which one might expect a high degree of correlation between measurements taken close together. In the case of elevations, a simplistic interpolation function takes the elevation value of the nearest stored location and assigns this to the location that is not stored. Smarter interpolation functions, involving more than a single stored value, should be preferred.

spatial autocorrelation

Tobler's first law of geography

Line objects, either by themselves or in their role as region object boundaries, are continuous phenomena that must be finitely represented. In real life, these objects are usually not straight, and can be erratically curved. A famous paradoxical question is whether one can actually measure the length of Great Britain's coastline, i.e. can one measure around rocks, pebbles or even grains of sand? In a computer, such random, curvilinear features can never be fully represented: they require a degree of generalization. Phenomena with intrinsic continuous and/or infinite characteristics therefore have to be represented with finite means (computer memory) for computer manipulation, yet any finite representation scheme is open to errors of interpretation. To allow for this, fields are usually implemented with a tessellation approach, and objects with a (topological) vector approach. In the following subsections we discuss tessellations and vector-based representations and how these are applied to represent geographic fields and objects.

Regular tessellations

Tessellation (or tiling) is a partitioning of space into mutually exclusive cells that together make up the complete study space. For each cell, a (thematic) value is assigned to characterize that part of space. Three types of regular tessellation are illustrated in Figure 8.4. In a regular tessellation, the cells have the same shape and size; a simple example of this is a rectangular raster of unit squares, represented in a computer in the 2D case as an array of $n \times m$ elements (see Figure 8.4 left).

Figure 8.4
The three most common types of regular tessellation: from left to right, square cells, hexagonal cells and triangular cells.
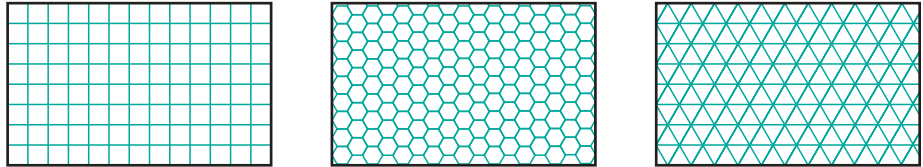
All regular tessellations have in common that the cells have the same shape and size, and that the field attribute value assigned to a cell is associated with the entire area occupied by the cell. Square cell tessellation is commonly used, mainly because geo-referencing of such a cell is straightforward. This type of tessellation is known under various names in different GIS packages: e.g. "raster" or "raster map".

grid

The size of the area that a single raster cell represents is called the raster's resolution. Sometimes, the term "grid" is used, but strictly speaking a grid is an equally spaced collection of points, all of which have some attribute value assigned. Grids are often used for discrete measurements that occur at regular intervals. Grid points are then considered synonymous to raster cells.

There are some issues related to cell-based partitioning of the study space. The field value of a cell can be interpreted as one for the complete tessellation cell, in which case the field is discrete, not continuous or even differentiable. Some convention is needed to state which value prevails on cell boundaries. With square cells, this convention states that lower and left boundaries belong to the cell. There are two approaches to refining the solution of this continuity issue: make the cell size smaller, so as to make the "continuity gaps" between the cells smaller; and/or assume that a cell value only represents elevation for one specific location in the cell, and to provide a good interpolation function for all other locations that have the continuity characteristic. If one wants to use rasters for continuous field representation, one usually uses the first approach but not the second, as the second technique is usually considered computationally too intensive for large rasters.

The location associated with a raster cell is fixed by convention: it may be the cell centroid (mid-point) or, for instance, its left lower corner. Values for other positions are computed using an interpolation function applied to one or more nearby field values. This allows us to represent continuous, even differentiable, functions. An important advantage of regular tessellations is that we know how they partition space, and that we can make our computations specific to this partitioning. This leads to fast algorithms. An obvious disadvantage is that they are not adaptive to the spatial phenomenon we want to represent. The cell boundaries are both artificial and fixed: they may or may not coincide with the boundaries of the phenomena of interest. If we use any of the above regular tessellations to represent an area with minor elevation differences, then, clearly we would need just as many cells as in a strongly undulating terrain: the data structure does not adapt to the lack of relief. We would, for instance, still use the $m \times n$ cells for the raster, even though variations in elevation are irrelevant.

Irregular tessellations

Regular tessellations provide simple structures with straightforward algorithms that are, however, not adaptive to the phenomena they represent. This means they might not represent the phenomena in the most efficient way. For this reason, substantial research effort has been put into irregular tessellation. Again, these are partitions of space into mutually distinct cells, but now the cells may vary in size and shape, allowing them to adapt to the spatial phenomena that they represent. Irregular tessellations are more complex than regular ones, but they are also more adaptive, which typically leads to a reduction in the amount of computer memory needed to store the data.
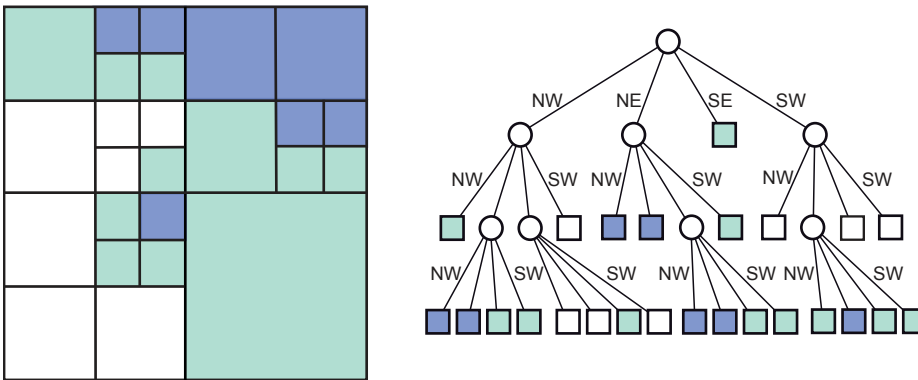
Figure 8.5
An 8×8, three-value raster (here, three colours) and its representation as a region quadtree. To construct a quadtree, the field is successively split into four quadrants until all parts have only a single value. After the first split, the southeast quadrant is entirely green, and this is indicated by a green square at level two of the tree. Other quadrants have to be split further.

quadtrees

A well-known data structure in this family—upon which many more variations have been based—is the region quadtree. It is based on a regular tessellation of square cells, but takes advantage of cases where neighbouring cells have the same field value, so that they can be represented together as one bigger cell. A simple illustration is provided in Figure 8.5. It shows a small 8×8 raster with three possible field values: white, green and blue. The quadtree that represents this raster is constructed by repeatedly splitting up the area into four quadrants, which are called NW, NE, SE, SW for obvious reasons. This procedure stops when all the cells in a quadrant have the same field value. The procedure produces an upside-down, tree-like structure, hence the name "quadtree". In the computer's main memory, the nodes of a quadtree (both circles and squares in Figure 8.5) are represented as records. The links between them are pointers, i.e. a programming technique to address (or to point to) other records. Quadtrees are adaptive because they apply Tobler's law (see Subsection 8.1.2). When a conglomerate of cells has the same value, they are represented together in the quadtree, provided their boundaries coincide with the predefined quadrant boundaries. Therefore, a quadtree provides a nested tessellation: quadrants are only split if they have two or more different values.

Vector representations

Tessellations do not explicitly store georeferences of the phenomena they represent. A georeference is a coordinate pair from some geographic space, and is also known as a vector. Instead, tesselations provide a georeference of the lower left corner of the raster, for instance, plus an indicator of the raster's resolution, thereby implicitly providing georeferences for all cells in the raster. In vector representations, georeferences are explicitly associated with the geographic phenomena. Examples and their vector representation are discussed below. To start, we will examine the TIN representation for geographic fields, which is a hybrid between tessellations and vector representations.

Triangulated Irregular Networks   A commonly-used data structure in GIS software is the triangulated irregular network, or TINs. It is a standard implementation techniques for digital terrain models, but it can also be used to represent any continuous field. The principles on which a TIN is based are simple. It is built from a set of locations for which we have a measurement, for instance an elevation. The locations can be arbitrarily scattered in space and are not usually on a regular grid. Any location together with its elevation value can be viewed as a point in three-dimensional space (Figure 8.6). From these 3D points, we can construct an irregular tessellation made of triangles. Two such tessellations are illustrated in Figure 8.7.

Figure 8.6
Input locations and their (elevation) values for a TIN construction. The location P is an arbitrary location that has no associated elevation measurement.
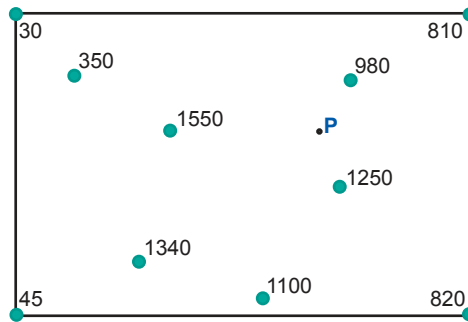
In three-dimensional space, three points uniquely determine a plane, as long as they are not collinear, i.e. they must not be positioned on the same line. A plane fitted through these points has a fixed aspect and gradient and can be used to compute an approximation of elevation of other locations. Since we can pick many triples of points, we can construct many such planes, and therefore we can have many elevation approximations for a single location, such as P (Figure 8.6).

By restricting the use of a plane to the triangular area "between" the three anchor points, we obtain a triangular tessellation of the complete study space. Unfortunately there are many different tessellations for a given input set of anchor points, as Figure 8.7 shows. Some tessellations are better than others, in the sense that they give smaller errors of elevation approximation. For instance, if we base our elevation computation for location P on the shaded triangle in the left-hand diagram, we will get a different value than that from the shaded triangle in the right-hand diagram. The latter will provide a better approximation because the average distance from P to the three triangle anchors is smaller.

Figure 8.7
Two triangulations based on the input locations of Figure 8.6: (a) one with many "stretched" triangles; (b) the triangles are more "equilateral", known as *Delaunay triangulation*.
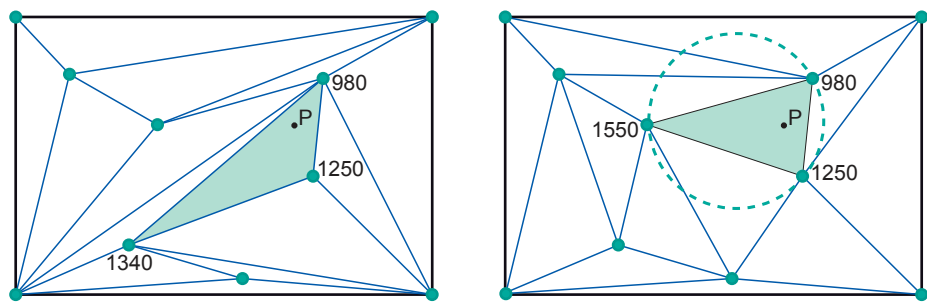
The triangulation of Figure 8.7b is a Delaunay triangulation, which in a sense is an optimal triangulation. There are several ways of defining such a triangulation (see [95]). Two important properties are, first, that the triangles are as equilateral ('equal-sided') as they can be, given the set of anchor points and, second, that for each triangle, the

circumcircle through its three anchor points does not contain any other anchor point. One such circumcircle is depicted on the right of Figure 8.7b. A TIN clearly is a vector representation: each anchor point has a stored georeference. Yet we might also call it an irregular tessellation, as the chosen triangulation provides a partitioning of the entire study space. However, in this case, the cells do not have an associated stored value as is typical of tessellations, but rather a simple interpolation function that uses the elevation values of its three anchor points.

Point representations   Points are defined as single coordinate pairs (x, y) in 2D space, or coordinate triplets (x, y, z) in 3D space. Points are used to represent objects that are best described as shapeless, size-less, zero-dimensional features. Whether this is the case really depends on the purposes of the application and also on the spatial extent of the objects compared to the scale used in the application. For a tourist map of a city, a park would not usually be considered a point feature, but perhaps a museum would, and certainly a public phone booth might be represented as a point. In addition to the georeference, administrative or thematic data are usually stored for each point object that can capture relevant information about it. For phone-booth objects, for example, this may include the telephone company owning the booth, its phone number and the date it was last serviced.

Line representations   Line data are used to represent one-dimensional objects such as roads, railroads, canals, rivers and power lines. Again, there is an issue of relevance for the application and the scale that the application requires. For the example of mapping tourist information, bus, subway and tram routes are likely to be relevant line features. Some cadastral systems, on the other hand, may consider roads to be two-dimensional features, i.e. having a width as well as length. Previously, we noted that arbitrary, continuous curvilinear features are as equally difficult to represent as continuous fields. GISs, therefore, approximate such features (finitely!) as lists of nodes: the two end nodes and zero or more internal nodes, or vertices, define a line. Other terms for "line" that are commonly used in some GISs are polyline, arc or edge. A node or vertex is like a point, but it only serves to define the line and provide shape in order to obtain a better approximation of the actual feature. The straight parts of a line between two consecutive vertices or end nodes are called line segments. Many GISs store a line as a sequence of coordinates of its end nodes and vertices, assuming that all its segments are straight. This is usually good enough, as cases in which a single straight line segment is considered an unsatisfactory representation can be dealt with by using multiple (smaller) line segments, instead of one.

arc, edge, node, vertex

Still, in some cases we would like to have the opportunity to use arbitrary curvilinear features to represent real-world phenomena. Think of a garden design with perfectly circular or elliptical lawns, or of detailed topographic maps showing roundabouts and the sidewalks. In principle all of this can be stored in a GIS, but currently many systems do not accommodate such shapes. A GIS function supporting curvilinear features uses parameterized mathematical descriptions, a discussion of which is beyond the scope of this textbook. Collections of (connected) lines may represent phenomena that are best viewed as networks. With networks, interesting questions arise that have to do with connectivity and network capacity. These relate to applications such as traffic monitoring and watershed management. With network elements—i.e. the lines that make up the network—extra values are commonly associated, such as distance, quality of the link or the carrying capacity.
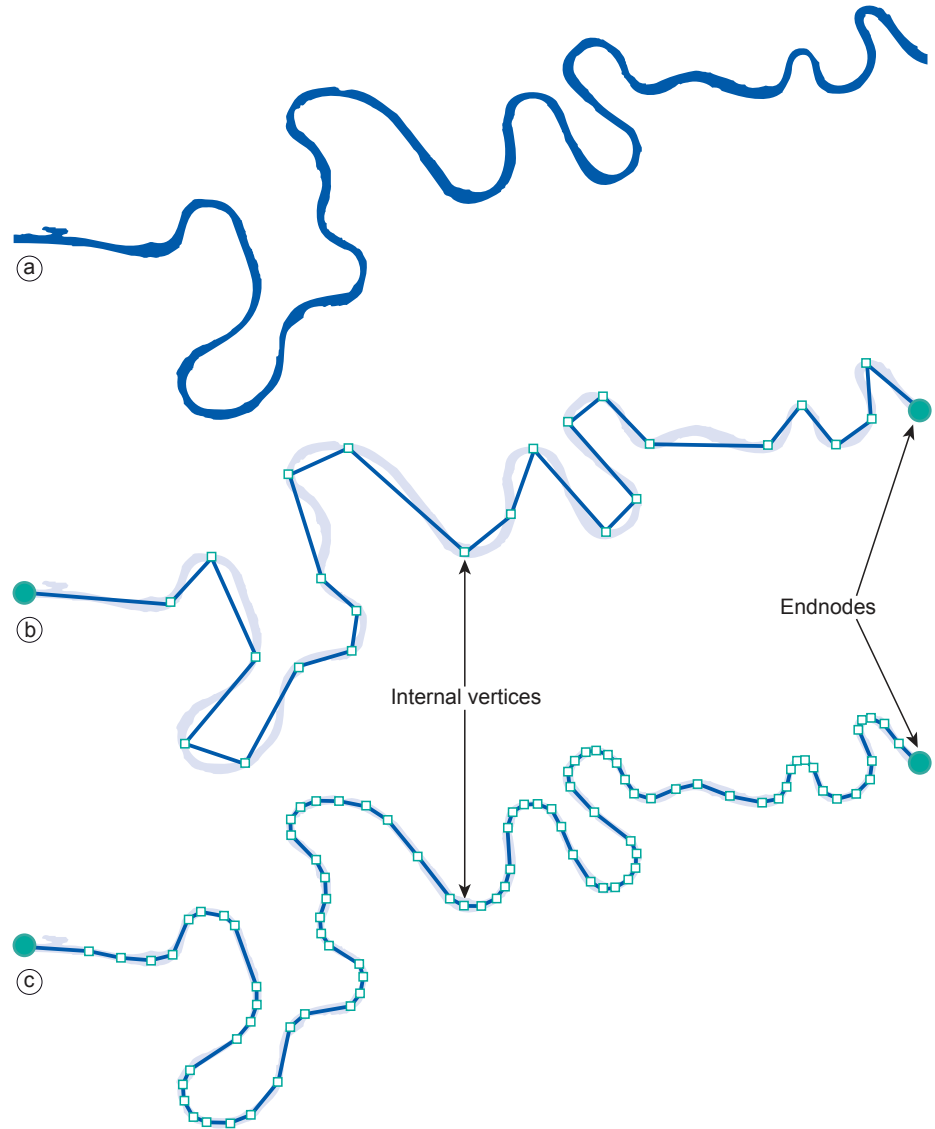
Figure 8.8
Examples of line
representation: (a) the
centerline of a river can be
represented as a line feature;
(b) this line feature consists
of a start node, an end node
and internal vertices; (c) by
increasing the number of
internal vertices, we can
closer resemble the shape of
the feature.

polygons

Area representations   When area objects are stored using a vector approach, the usual technique is to apply a boundary model. This means that each area feature is represented by some arc/node structure that determines a polygon as the area's boundary. Common sense dictates that area features of the same kind are best stored in a single data layer, represented by mutually non-overlapping polygons. This results in an application-determined (i.e. adaptive) partition of space. A polygon representation for an area object is another example of a finite approximation of a phenomenon that may have a curvilinear boundary in reality. If the object has a fuzzy boundary, a polygon is an even worse approximation, even though potentially it may be the only one possible. Figure 8.9 illustrates a simple study with three area objects, each represented by polygon boundaries. Clearly, we expect additional data to accompany the area data. Such information could be stored in database tables.

A simple but naïve representation of area features would be to list for each polygon

the list of lines that describes its boundary. Each line in the list would, as before, be a sequence that starts with a node and ends with one, possibly with vertices in between. A closer look at the shared boundary between the bottom left and right polygons in Figure 8.9 shows why this approach is far from optimal. As the same line makes up the boundary from the two polygons, this line would be stored twice in the above representation, namely once for each polygon. This is a form of data duplication—known as data redundancy—which is (at least in theory) unnecessary, although it remains a feature of some systems. Another disadvantage of such polygon-by-polygon representations is that if we want to identify the polygons that border the bottom left polygon, we have to do a complicated and time-consuming search analysis comparing the vertex lists of all boundary lines with that of the bottom left polygon. For Figure 8.9, with just three polygons, this is fine, but in a data set with 5000 polygons, and perhaps a total of 25,000 boundary lines, this becomes a tedious task, even with the fastest of computers.
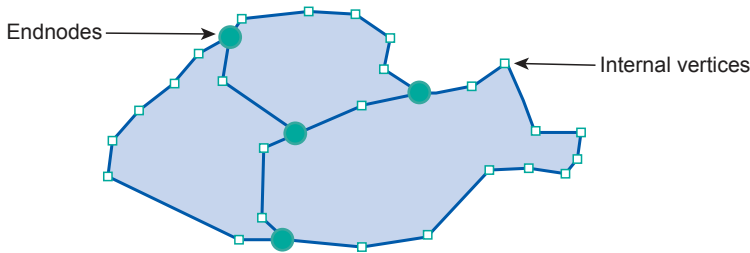
The boundary model is an improved representation that deals with these disadvantages. It stores parts of a polygon's boundary as non-looping arcs and indicates which polygon is on the left and which is on the right of each arc. A simple example of the boundary model can be seen in Figure 8.10. It illustrates which additional information is stored about spatial relationships between lines and polygons. Obviously, real coordinates for nodes (and vertices) will also be stored in another table. The boundary model is also called the topological data model as it captures some topological information, such as polygon neighbourhood, for example. Observe that it is a matter of a simple query to find all the polygons that are the neighbour of a given polygon, unlike the case above.

line	from	to	left	right	vertexlist
b_1	4	1	W	A	...
b_2	1	2	B	A	...
b_3	1	3	W	B	...
b_4	2	4	C	A	...
b_5	3	4	W	C	...
b_6	3	2	C	B	...
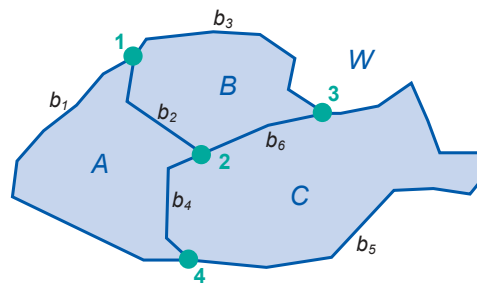
Topology and spatial relationships

General spatial topology Topology deals with spatial properties that do not change under specific transformations. Take features (as in Figure 8.11) drawn on a sheet of rubber. These features can be made to change in shape and size by stretching and pulling the sheet, yet some properties of these features will not change:

- area E is still inside area D;

- the neighbourhood relationships between A, B, C, D, and E stay intact, and their boundaries have the same start and end nodes;

- the areas are still bounded by the same boundaries, only the shapes and lengths of their perimeters have changed.

topological relationships

Topological relationships are built from simple elements into more complex elements: nodes define line segments, and line segments connect to define lines, which in turn define polygons. Issues relating to order, connectivity and adjacency of geographical elements form the basis of more sophisticated GIS analyses. These relationships (called topological properties) are invariant under a continuous transformation and are referred to as a topological mapping.

We will now consider topological aspects in two ways. Firstly, using simplices, we will look at how simple elements define more complex ones. Secondly, we will examine the logical aspects of topological relationships using set theory. The three-dimensional case is also briefly discussed.
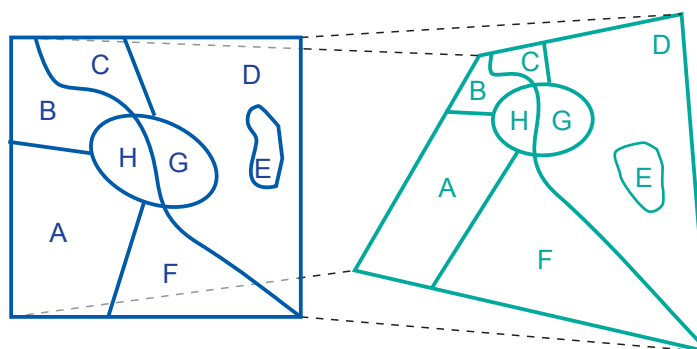
Figure 8.11
Rubber sheet transformation: the space is transformed, yet many relationships between the constituents remain unchanged.

Topological relationships   The mathematical properties of the geometric space used for spatial data may be described as follows:

- The space is a 3D Euclidean space in which we can determine for every point its coordinates as a triple (x, y, z) of real numbers. In this space, we can define features such as points, lines, polygons and volumes as geometric primitives of the respective dimension. A point is a zero-dimensional, a line a one-dimensional, a polygon a two-dimensional, and a volume a three-dimensional primitive.

- The space is a metric space, which means that we can always compute the distance between two points according to a given distance function. Such a function is also known as a metric.

- The space is a topological space, the definition of which is a bit complicated. In essence, for every point in the space we can find a neighbourhood around it that fully belongs to that space as well.

- Interiors and boundaries are properties of spatial features that remain invariant under topological mappings. This means that, under any topological mapping, the interior and the boundary of a feature remains unbroken and intact.

A number of advantages exist when our computer representations of geographic phenomena have built-in sensitivity to topological issues. Questions related to the "neighbourhood" of an area are a case in point. To obtain some "topological sensitivity", simple building blocks have been proposed with which more complicated representations can be constructed:

- We can define features within the topological space that are easy to handle and that can be used as representations of geographic objects. These features are called simplices as they are the simplest geometric shapes of some dimension:
  - point (0-simplex),
  - line segment (1-simplex),
  - triangle (2-simplex),
  - and tetrahedron (3-simplex).

- When we combine various simplices into a single feature, we obtain a simplicial complex; see Figure 8.12 for examples.

As the topological characteristics of simplices are well-known, we can infer the topological characteristics of a simplicial complex from the way it was constructed.
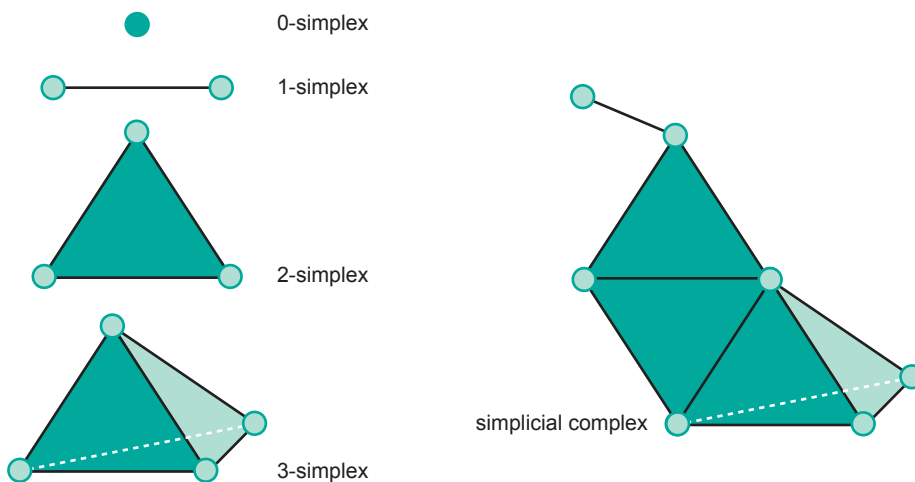
Figure 8.12
Simplices and a simplicial complex. Features are approximated by a set of points, line segments, triangles and tetrahedrons.

The topology of two dimensions   We can use the topological properties of interiors and boundaries to define relationships between spatial features. Since the properties of interiors and boundaries do not change under topological mapping, we can investigate their possible relations between spatial features. We can define the *interior* of a region, R, as the largest set of points of R for which we can construct a disc-like environment around it (no matter how small) that also falls completely inside R. The boundary of R is the set of those points belonging to R that do not belong to the interior of R, i.e. one cannot construct a disc-like environment around such points that still belongs to R completely.

interior and exterior

Let us consider a spatial region A. It has a boundary and an interior, both seen as (infinite) sets of points, which are denoted by *boundary*(A) and *interior*(A), respectively. We consider all possible combinations of intersections ($\cap$) between the boundary and

the interior of A with those of another region, B, and test whether they are the empty set ($\emptyset$) or not. From these intersection patterns, we can derive eight (mutually exclusive) spatial relationships between two regions. If, for instance, the interiors of A and B do not intersect, but their boundaries do, yet the boundary of one does not intersect the interior of the other, we say that A and B meet. In mathematics, we can therefore define the "meets relationship" using set theory as:

$$A \; meets \; B \quad \stackrel{\text{def}}{=} \quad \begin{aligned} & interior(A) \cap interior(B) = \emptyset \;\wedge \\ & boundary(A) \cap boundary(B) \neq \emptyset \;\wedge \\ & interior(A) \cap boundary(B) = \emptyset \;\wedge \\ & boundary(A) \cap interior(B) = \emptyset. \end{aligned}$$

In the above formula, the symbol $\wedge$ expresses the logical connective "and". Thus, the formula states four properties that must all hold for the formula to be true.

Figure 8.13 shows all eight spatial relationships: *disjoint*, *meets*, *equals*, *inside*, *covered by*, *contains*, *covers* and *overlaps*. These relationships can be used, for instance, in queries on a spatial database. Rules of how simplices and simplicial complexes can be embedded in 2D and 3D space are quite different. Such a set of rules defines the topological consistency of that space. It can be proven that if the rules presented and illustrated in Figure 8.14 are satisfied for all features in 2D space, then the features define a topologically consistent configuration in 2D space.

topological consistency
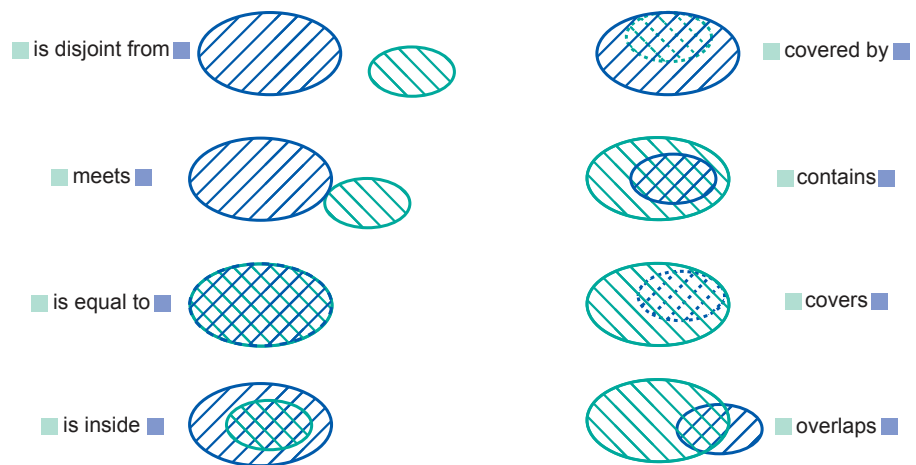
Figure 8.13
Spatial relationships between two regions derived from the topological invariants of intersections of boundary and interior.

The topology of three dimensions   Our discussion of vector representations and spatial topology has so far focused on objects in 2D space. The history of spatial data handling is almost purely 2D, and this is remains the case for the majority of present-day GIS applications. Many application domains make use of elevational data, but these are usually accommodated for by what are known as 2.5D data structures. These 2.5D data structures are similar to the 2D data structures just discussed, using points, lines and areas. They also apply the rules of two-dimensional topology, as illustrated in Figure 8.14. This means that different lines cannot cross without intersecting nodes and that different areas cannot overlap. There is, however, one important aspect in

1. Every 1-simplex ('arc') must be bounded by two 0-simplices ('nodes', namely its begin and end node).
2. Every 1-simplex borders two 2-simplices ('polygons', namely its 'left' and 'right' polygons).
3. Every 2 -simplex has a closed boundary consisting of an alternating (and cyclic) sequence of 0- and 1-simplices.
4. Around every 0-simplex exists an alterning (and cyclic) sequence of 1- and 2-simplices.
5. 1-simplices only intersect at their (bounding) nodes.

Rule (1)   Rule (2)   Rule (3)   Rule (4)   Rule (5)
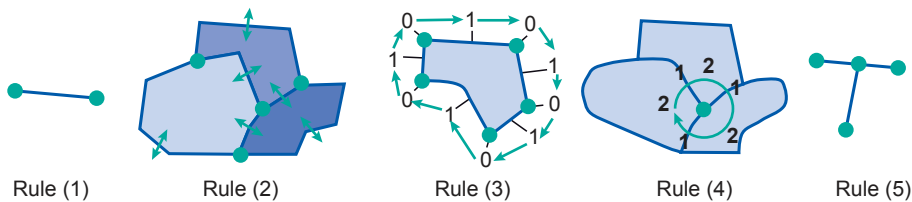
Figure 8.14
The five rules of topological consistency in two-dimensional space.

which 2.5D data do differ from standard 2D data and that is in their association of an additional $z-$value with each 0-simplex ('node'). Thus, nodes also have an elevation value associated with them. Essentially, this allows the GIS user to represent 1- and 2-simplices that are non-horizontal, so that a piece-wise planar, "wrinkled surface" can be constructed as well, much like a TIN. One cannot have two different nodes with identical x and y coordinates but different z values. Such nodes would constitute a perfectly vertical feature, which is not allowed. Consequently, true solids cannot be represented in a 2.5D GIS.

Solid representation is an important feature for some dedicated GIS application domains. Two examples of such applications are: mineral exploration, where solids represent ore bodies; and urban models, where solids represent various human constructions, such as buildings and sewers. The 3D characteristics of such objects are fundamentally important since their depth and volume may matter, or their real life visibility must be faithfully represented. A solid can be defined as a true 3D object. An important class of solids in 3D GISs is formed by polyhedra, which are the solids limited by their planar facets. A facet is polygon-shaped, flat side that is part of the boundary of a polyhedron. Any polyhedron has at least four facets, which happens to be the case for the 3-simplex. Most polyhedra have many more facets, e.g. a cube already has six.

Scale and resolution

In the practice of spatial data handling, one often comes across questions like "What is the resolution of the data?" or "At what scale is your data set?" Now that we have moved firmly into the digital age, these questions sometimes defy an easy answer. Map scale can be defined as the ratio between the distance on a printed map and the distance of the same stretch in the terrain. A 1:50,000 scale map means that 1 cm on the map represents 50,000 cm (i.e. 500 m) in the terrain. "Large-scale" means that the ratio is relatively large, so typically it means there is much detail to see, as on a 1:1000 printed map. "Small-scale", in contrast, means a small ratio, hence less detail, as on a 1:2,500,000 printed map. When applied to spatial data, the term resolution is commonly associated with the cell width of the tessellation applied.

large-scale and small-scale maps

Digital spatial data, as stored in a GIS, are essentially without scale: scale is a ratio notion associated with visual output, such as a map or on-screen display, not with the data that was used to produce the map or display. When digital spatial data sets have been collected with a specific map-making purpose in mind, and all maps have been designed to use one single map scale, for instance 1:25,000, we may assume that the

data carries the characteristic of "a 1:25,000 digital data set."

Representations of geographic fields

We have looked at various representation techniques in some detail. Now we can study which of them can be used to represent a geographic field. A geographic field can be represented by means of a tessellation, a TIN or a vector representation. The choice between them is determined by the requirements of the application in mind. It is more common to use tessellations, notably rasters, for field representation, but vector representations are in use too. We have already looked at TINs, so the following subsections only present examples of the other two representation techniques.

Raster representation of a field   Figure 8.15 illustrates how a raster represents a continuous field, in this case elevation. Different shades of blue indicate different elevation values, with darker blue tones indicating higher elevations. The choice of a blue spectrum is only to make the illustration aesthetically pleasing; real elevation values are stored in the raster, so we could have printed a real number value in each cell instead. This would not have made the figure very legible, however. A raster can be thought of as a long list of field values: actually, there should be $m \times n$ such values present. The list is preceded with some extra information, such as a single georeference for the origin of the whole raster, a cell-size indicator, the integer values for m and n, and an indicator of data type for interpreting cell values. Rasters and quadtrees do not store the georeference of each cell, but infer it from the extra information about the raster. A TIN is a much "sparser" data structure: as compared to a regular raster, the amount of data stored is less for a structure of approximately equal interpolation error. The quality of the TIN depends on the choice of anchor points, as well as on the triangulation built from it. It is, for instance, wise to perform "ridge following" during the data acquisition process for a TIN. Anchor points on elevation ridges will assist in correctly representing peaks and faces of mountain slopes.

Figure 8.15
A raster representation (in part) of the elevation of the study area of Figure 8.1 (Falset, Spain). Actual elevation values are indicated in shades of blue. The depicted area is the northeast flank of the mountain in the southeastern part of the study area. The right-hand figure zooms in on a part of the left-hand figure.
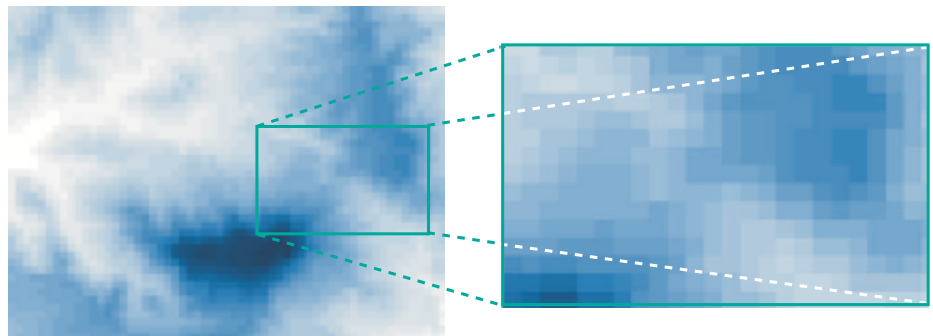
Vector representation of a field   We briefly mention the vector representation for fields such as elevation, which uses isolines of the field. An isoline is a linear feature that connects points with equal field values. When the field is elevation, we also speak of contour lines. Elevations in the Falset study area are represented by contour lines in Figure 8.16. Both TINs and isoline representations use vectors. Isolines as a representation mechanism are not common, however. They are used as a geoinformation visualization technique (in mapping, for instance), but usually it is better to choose a TIN for representing this type of field. Many GIS packages provide functions to generate an isoline visualization from a TIN.

Representation of geographic objects

The representation of geographic objects is most naturally supported with vectors. After all, objects are identified by the parameters of location, shape, size and orientation (see Section 8.1.1), and many of these parameters can be expressed in terms of vectors. Tessellations are also commonly used for representing geographic objects.

Tessellations for representing geographic objects   Remotely-sensed images are an important data source for GIS applications. Unprocessed digital images contain many pixels, each of which carrying a reflectance value. Various techniques exist to process digital images into classified images that can be stored in a GIS as a raster. Image classification characterizes each pixel into one of a finite list of classes, thereby obtaining an interpretation of the contents of the image. The recognized classes can be crop types, as in the case of Figure 8.17, or urban land use classes, as in the case of Figure 8.18. These figures illustrate the unprocessed images (a) and a classified version of the image (b). For the application at hand, perhaps only potato fields (Figure 8.17b, in yellow) or industrial complexes (Figure 8.18b, in orange) are of interest. This would mean that all other classes are considered unimportant and would probably be ignored in further analysis. If that further analysis can be carried out with raster data formats, then there is no need to consider vector representations.

Nonetheless, we must make a few observations regarding the representation of geographic objects in rasters. Line and point objects are more awkward to represent using rasters. Area objects, however, are conveniently represented in a raster, although area boundaries may appear as jagged edges. This is a typical by-product of raster resolution versus area size, and artificial cell boundaries. This may have consequences for area-size computations: the precision with which the raster defines the object's size is limited. After all, we could say that rasters are area based and that geographic objects

that are perceived as lines or points are considered to have zero area size. Standard classification techniques may, moreover, fail to recognize these objects as points or lines.
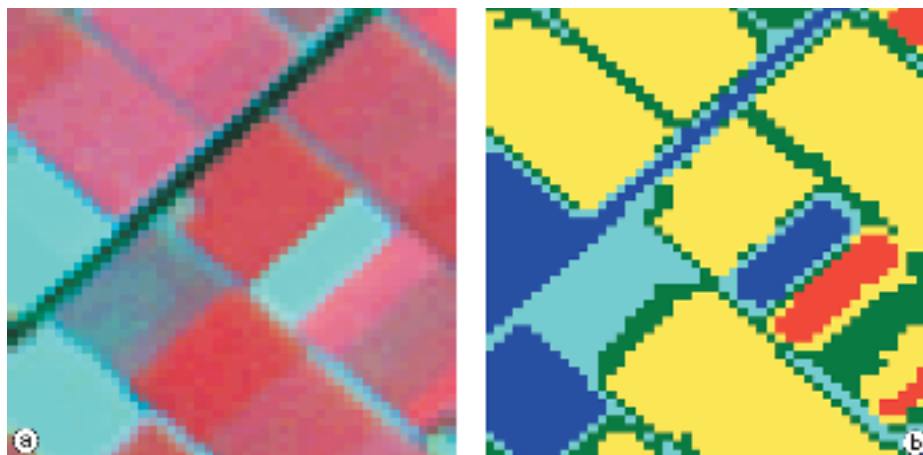
Figure 8.17
An unprocessed digital image (a) and a classified raster (b) of an agricultural area.
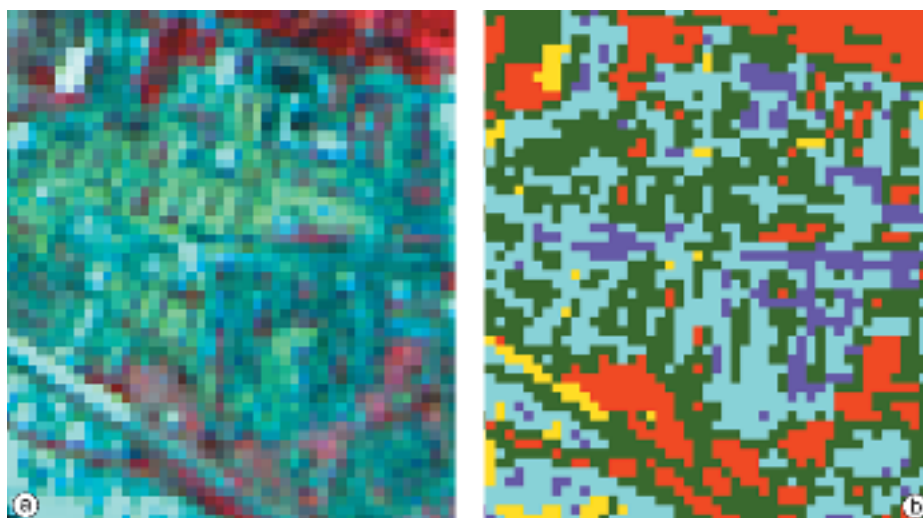
Figure 8.18
An unprocessed digital image (a) and a classified raster (b) of an urban area.

Many GISs support line representations in a raster, as well as operations on them. Lines can be represented as strings of neighbouring raster cells of equal value (Figure 8.19). Supported operations are connectivity operations and distance computations. Note that the issue of the precision of such computations needs to be addressed.

Vector representations of geographic objects   A more natural way of depicting geographic objects is by means of vector representations. Most of the issues related to this have already been discussed in Subsection 8.1.2, so a small example should suffice here. Figure 8.20 shows a number of geographic objects in the vicinity of the ITC building. These objects are area representations in a boundary model. Nodes and vertices of the polylines that make up the objects' boundaries are not illustrated, though obviously they have been stored.
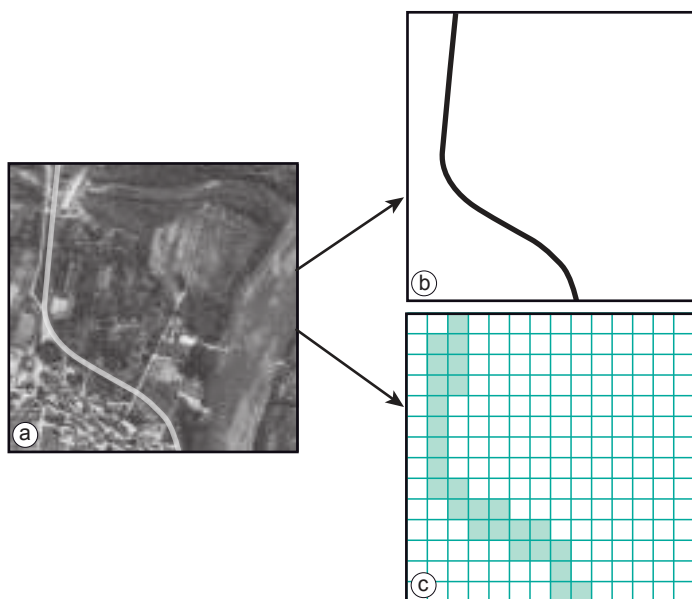
Figure 8.19
A linear feature (a) represented as a vector line feature (b) and in a raster layer (c).
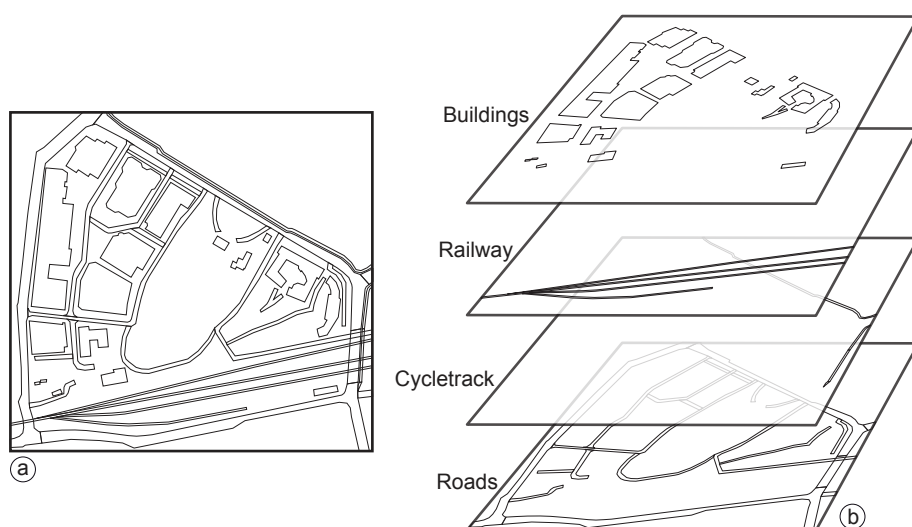
Figure 8.20
Various objects displayed as area objects in a vector representation. Similar data types are stored in the same single layer (e.g. Buildings). For each different type a new layer is used (b).

8.1.3 Organizing and managing spatial data

In Subsection 8.1.2 we discussed various types of geographic information and ways of representing it. We did not, however, pay much attention to how various sorts of spatial data can be combined in a single system.

The main principle of data organization applied in a GIS is that of spatial data layers. A spatial data layer is either a representation of a continuous or discrete field, or a collection of objects of the same kind. Usually, the data are organized such that similar elements are in a single data layer. For example, all telephone booth point objects would be in one layer, and all road line objects in another. A data layer contains spatial data—of any of the types discussed above—as well as attribute (i.e. thematic) data, which further describes the field or objects in the layer. Attribute data are quite

often arranged in tabular form, maintained in some kind of geo-database, as we will see in Section 8.4. An example of two field-data layers is provided in Figure 8.21.
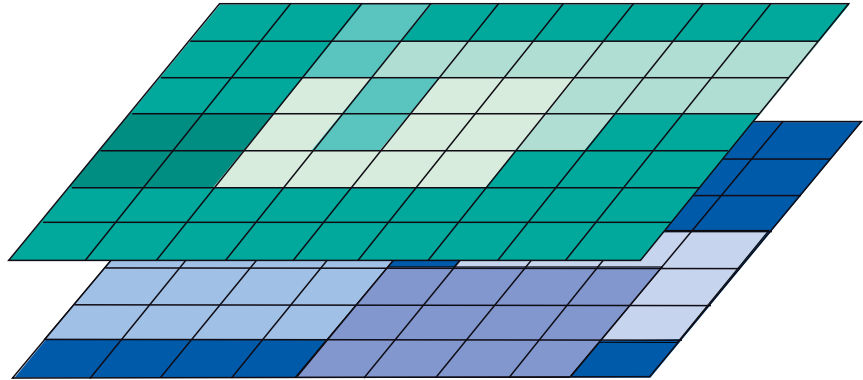
Figure 8.21
Different rasters can be overlaid to look for spatial correlations.

Data layers can be laid over each other, inside a GIS package, to study combinations of geographic phenomena. We shall see below that a GIS can be used to study the spatial correlation between different phenomena, albeit requiring computations that overlay one data layer with another. This is schematically depicted in Figure 8.22 for two different object layers. Field layers can also be involved in overlay operations. For a more detailed discussion of the functions offered for data management by GISs and database systems, refer to Chapter 9.
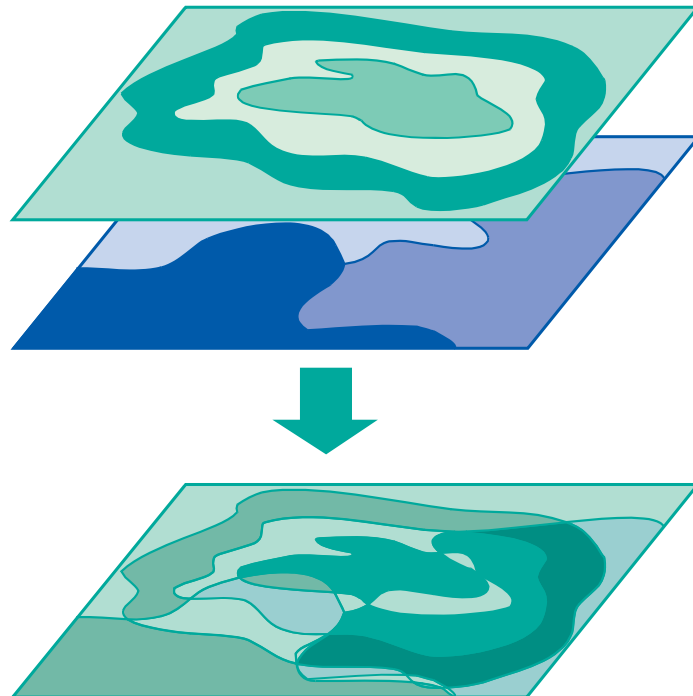
Figure 8.22
Two different object layers can be overlaid to look for spatial correlations; the result can be used as a separate (object) layer.

8.1.4 The temporal dimension

Besides having geometric, thematic and topological properties, geographic phenomena also change over time and are thus dynamic. Examples include identifying the

owners of a land parcel in 1972, or determining how land cover in a certain area changed from native forest to pasture land over a specific time period. Some features or phenomena change slowly, e.g. geological features or land cover, as in the example just given. Other phenomena change very rapidly, such as the movement of people or atmospheric conditions. Some examples are provided in Figure 8.23. For an increasing number of applications, these changes themselves are the key aspect of the phenomenon to be studied. For different applications, different scales of measurement will apply.

dynamic phenomena

- Where and when did something happen?

- How fast did this change occur?

- In which order did the changes occur?

The way we represent relevant components of the real world in our models determines the kinds of questions we can or cannot answer. In this chapter we have already discussed representation issues for spatial features, but so far we have ignored issues for incorporating time. The main reason is that GISs still offer limited support for doing so. As a result, most studies require substantial efforts from the GIS user in data preparation and data manipulation. Also, besides representing an object or field in 2D or 3D space, the temporal dimension is of a continuous nature. Therefore, in order to represent it in a GIS we have to discretize the time dimension.

time in GIS

Spatio-temporal data models are ways of organizing representations of space and time in a GIS. Several representation techniques have been proposed in the literature. Perhaps the most common of these is the "snapshot state", which represents a single moment in time of an ongoing natural or man-made process. We may store a series of these "snapshot states" to represent "change", but we must be aware that this is by no means a comprehensive representation of that process. Further treatment of spatio-temporal data models is outside the scope of this book and readers are referred to Langran [64] for a discussion of relevant concepts and issues.

As time is the central concept of the temporal dimension, a brief examination of the nature of time may clarify our thinking when we work with this dimension:

- Discrete and continuous time: Time can be measured along a discrete or continuous scale. Discrete time is composed of discrete elements (seconds, minutes, hours, days, months, or years). For continuous time, no such discrete elements exist: for any two moments in time there is always another moment in between. We can also structure time by events (moments) or periods (intervals). When we represent intervals by a start and an end event, we can derive temporal relationships between events and periods, such as "before", "overlap", and "after".

- Valid time and transaction time: Valid time (or world time) is the time when an event really happened, or a string of events took place. Transaction time (or database time) is the time when the event was stored in the database or GIS. Note that the time at which we store something in a database is typically (much) later than when the related event took place.

- Linear, branching and cyclic time: Time can be considered to be linear, extending from the past to the present ('now'), and into the future. This view gives a single time line. For some types of temporal analysis, branching time—in which different time lines from a certain point in time onwards are possible—and cyclic time—in which repeating cycles such as seasons or days of the week are recognized—make more sense and can be useful.

- Time granularity: When measuring time, we speak of granularity as the precision of a time value in a GIS or database (e.g. year, month, day, second). Different applications may obviously require different granularity. In cadastral applications, time granularity might well be a day, as the law requires deeds to be date-marked; in geological mapping applications, time granularity is more likely to be in the order of thousands or millions of years.

- Absolute and relative time: Time can be represented as absolute or relative. Absolute time marks a point on the time line where events happen (e.g. "6 July 1999 at 11:15 p.m."). Relative time is indicated relative to other points in time (e.g. "yesterday", "last year", "tomorrow", which are all relative to "now", or "two weeks later", which is relative to some other arbitrary point in time.).
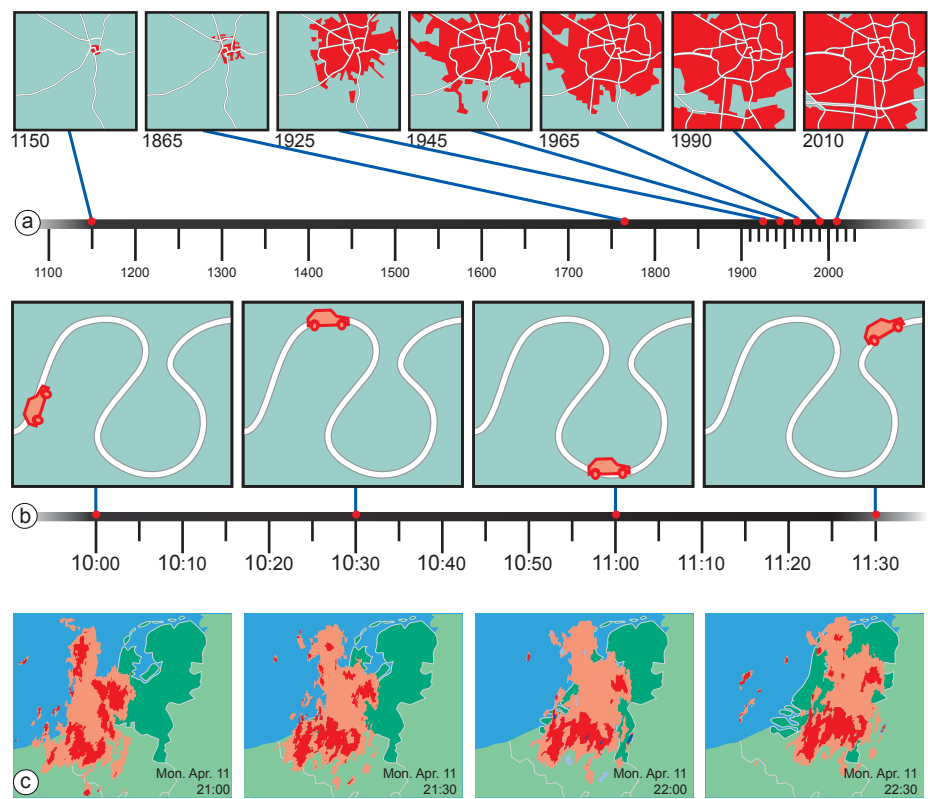
Figure 8.23
Examples of spatio-temporal phenomena: (a) expansion of a city, the area covered by the city grows over time, but the location of the city does not change; (b) a moving car will change position, but the object car does not change size or shape; (c) over time, the position of a cloud will change, but also the size and shape of the cloud can undergo changes over time.

In spatio-temporal analysis we consider changes of spatial and thematic attributes over time. We can keep the spatial domain fixed and look only at the attribute changes over time for a given location in space. We might be interested how land cover has changed for a given location or how land use has changed for a given land parcel over time, provided its boundary has not changed. On the other hand, we can keep the attribute domain fixed and consider the spatial changes over time for a given thematic attribute. In this case, we might want to identify locations that were covered by forest over a given period of time.

spatio-temporal analysis

Finally, we can assume both the spatial and attribute domains are variable and consider how fields or objects have changed over time. This may lead to notions of object motion—a subject receiving increasing attention in the literature. Applications

of moving object research include traffic control, mobile telephony, wildlife tracking, vector-borne disease control and weather forecasting. In these types of applications, the problem of object identity becomes apparent. When does a change or movement cause an object to disappear and become something new? With wildlife this is quite obvious; with weather systems less so. But this should no longer be a surprise: we have already seen that some geographic phenomena can be nicely described as objects, while others are better represented as fields.

8.2 Data entry

Spatial data can be obtained from various sources. It can be collected from scratch, using direct spatial-data acquisition techniques, or indirectly, by making use of existing spatial data collected by others. The first source could include field survey data and remotely sensed images. To the second source belongs printed maps and existing digital data sets. This section discusses the collection and use of data from both sources.

8.2.1 Spatial data input

One way to obtain spatial data is by direct observation of relevant geographic phenomena. This can be done through ground-based field surveys or by using remote sensors on satellites or aircraft. Many Earth science disciplines have developed specific survey techniques as ground-based approaches remain the most important source of reliable data in many cases.

primary data

Data that are captured directly from the environment are called *primary data*. With primary data, the core concern in knowing their properties is to know the process by which they were captured, the parameters of any instruments used, and the rigour with which quality requirements were observed.

In practice, it is not always feasible to obtain spatial data by direct capture. Factors of cost and available time may be a hindrance, and sometimes previous projects have acquired data that may fit a current project's purpose.

secondary data

In contrast to direct methods of data capture, spatial data can also be sourced indirectly. This includes data derived by scanning existing printed maps, data digitized from a satellite image, processed data purchased from data-capture firms or international agencies, and so on. This type of data is known as *secondary data*. Secondary data are derived from existing sources and have been collected for other purposes, often not connected with the investigation at hand.

Key sources of primary and secondary data, and several issues related to their use in analyses that users should be aware of, are discussed in the remainder of this section.

8.2.2 Aerial surveys and satellite remote sensing

Aerial photographs are a major source of digital data (see Section 4.6); soft-copy work-stations are used to digitize features directly from stereo pairs of digital photographs. These systems allow data to be captured in two or three dimensions, with elevations measured directly from a stereo pair using the principles of photogrammetry. Analogue aerial photos are often scanned before being entered into a soft-copy system, but with the advance of high-quality digital cameras this step can now be skipped.

In general, the alignment of roads and railways, lakes and water, and shapes of buildings are easily interpreted on aerial photographs—assuming that the scale of the photographs is not too small. Also, constructions such as dikes, bridges, air fields and the main types of vegetation and cultivation are mostly clearly visible. Nevertheless, numerous attribute data related to terrain features cannot be interpreted on aerial photographs: e.g. the administrative qualification of roads, sea and lake depths, functions of buildings, street names, and administrative boundaries. We will have to collect this information in the field or from existing data sets and maps (e.g. road maps, navigational charts or town plans). Issues related to the process of visual image interpretation are discussed in greater detail in Section 6.1.

Satellite remote sensing is another important source of spatial data. For this, satellites use different sensor packages to passively measure reflectance of parts of the elec-

tromagnetic spectrum or radio waves that were emitted by an active sensor such as radar (see Sections 4.4). Remote sensing collects raster data that can be further processed using different wavelength bands to identify objects and classes of interest, e.g. land cover. Issues related to the pre-processing of satellite remote-sensing data are discussed in greater detail in Chapter 5.

Figure 8.24
Aerial surveys (a) and satellite remote sensing (b) are employed to map relatively large areas at comparably large scales, source : Shuttle Radar Topography Mission, U.S. Geological Survey Department of the Interior/USGS and NASA, JPL.

8.2.3 Terrestrial surveys

Terrestrial surveys are usually employed for details that must be measured accurately, e.g. survey control stations, property boundaries, buildings, and road construction works (Figure 8.25). The surveyed data are often used to supplement and update existing data and for verification of data collected from aerial surveys and by satellite remote sensing. A terrestrial survey records features showing their relative position both horizontally and vertically. Several surveying techniques are employed to do this.

In horizontal positioning, measured angles at, and at distances from, known points are used to determine the positions of other points. Traditionally, survey measurements were made with optical and mechanical surveying instruments, such as a theodolite to measure angles, and more accurate electronic and electro-optical devices such as lasers for measuring distances. A more modern instrument is a total station, which is a theodolite with an electronic distance measurement device. Since the introduction of total stations, there has been a technological shift from the use of optical-mechanical devices to fully electronic systems incorporating a computer and relevant software.

Though satellite receivers are used for terrestrial surveying, total stations are still used widely, along with other types of surveying instruments, because of their accuracy, and area of operation: satellite systems do not work well in areas with dense tree cover or a high density of buildings.

Vertical positioning is usually done by levelling, which is a technique for measuring differences in height between established points relative to a datum or base point. Over short distances, levelling telescopes are used to view a staff or pole and, with the aid of a bar code, the height is recorded in relation to the previous station (Figure 8.25).

Elevation heights can also be derived with satellite receivers, albeit usually with some-
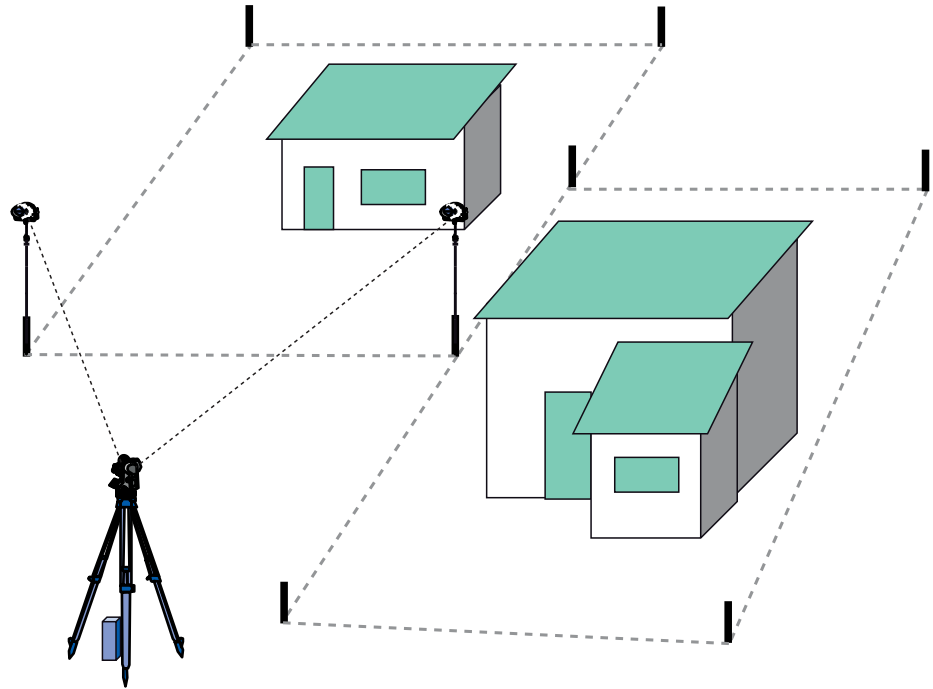
Figure 8.25
Terrestrial surveys are used
to map details such as
property boundaries.

Figure 8.26
Satellite-based surveys
enable efficient data
collection in open areas.

what less accuracy than for traditional precise levelling. However, the accuracy of satellite receivers may be similar if traditional levelling has to be used over a long distance. Laser altimetry (see Section 4.5) is employed for large areas, but its accuracy is not as good as levelling or GPS.

8.2.4 Field surveys

Every field science in natural resources, water resources, and urban and regional planning has a range of techniques for collecting data in the terrain. Full details of these techniques are given in various standard texts for the disciplines concerned.

Field surveys of natural and water resources are frequently carried out to check and supplement information derived from the interpretation of aerial photographs and satellite imagery (Figure 8.27). Often, sample areas are chosen within a study area for more detailed investigations. Socio-economic data, however, are often collected on the basis of administrative districts, with the result that their location is insufficiently precise to permit analysis of high quality.

Primary socio-economic data are collected by interviews and questionnaires. If the investigation is unofficial, the response will depend on the type of information required. In general, information of a financial or personal nature is difficult to come by and, even if given, may not be wholly reliable.

Fortunately a wealth of societal and economic data is available from official sources. Private individuals and commercial undertakings are usually required to provide government agencies with information via censuses, tax returns, etc. Since much of this data is confidential, it will usually be refined and generalized before it is released to others.

An example of a publicly available statistical data set is the International Data Base (IDB) provided by the US Census Bureau. It contains demographic and socio-economic statistics collected for 227 countries and areas of the world. The major types of data made available by the IDB are population by age and sex, birth and dead rates, migration, ethnicity, religion, language, literacy, labour force, employment, income and household composition.

Mobile GIS

Until recently, printed maps and forms were taken to the field and the information collected was sketched as notes on the map or written down on a form. This information was entered into a GIS database after returning to the office. This method of data collection is inefficient and prone to error. With a mobile GIS system and the support of a satellite receiver, we can take a GIS into the field with us on powerful, compact mobile computers and view, capture and update information, and then synchronize changes between the field and office (Figure 8.28).

Professional applications for mobile GISs are endless—utilities, forestry, environmental monitoring, field engineering, to mention a few. With the integration of systems, users are able to view each others' locations and, for example, share field data dynamically across their organization. Specifically, the data captured with mobile GISs can be instantly checked, updated and exchanged if necessary.

A simple task-driven mobile application begins in the office. GIS data are extracted from the main database and mapped onto the mobile device to be used in the field. The updated data are uploaded after returning to the office (Figure 8.29a).

A high-end mobile GIS application typically runs on a powerful laptop computer,
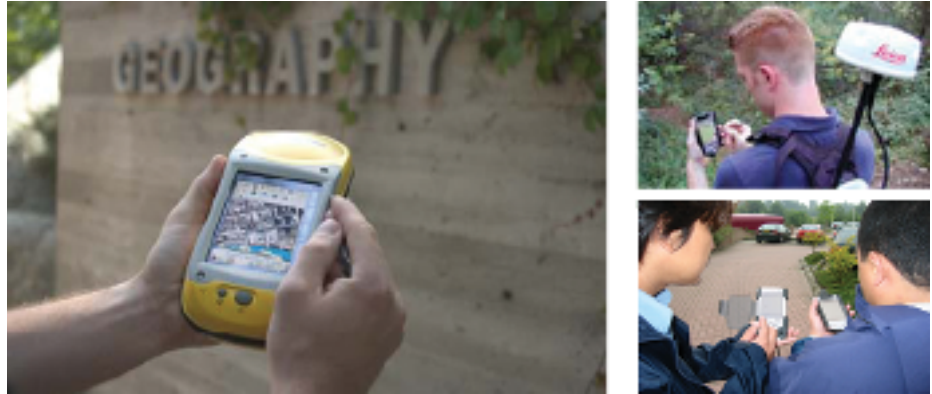
Figure 8.28
Mobile GIS provides the integration of mapping, GIS and positioning to field users via hand-held and mobile devices.

many of which provide a rich set of tools comparable to a desktop GIS application. A fast wireless connection enables direct access to maps and databases at the office, and synchronizes changes between the field and office through a web service (Figure 8.29b).

In cases where there is no connection to the main database in the office (e.g. a firewall makes access impossible), field edits can be synchronized later, when access to the main database is provided (Figure 8.29c). A versioned transaction may take care of the situation that the same feature (in the field) is updated several times: it can compare the updates (reconciling the version edits) before transferring the feature to the main, or parent, database.

Figure 8.29
Mobile updating strategies: (a) Data are extracted from the main (enterprise) database and mapped onto the mobile device. Field edits are uploaded to the main database after returning to the office. (b) Wireless connection between field and office enables real-time updating. (c) Multiple versions of the database are used to enable updating by multiple mobile users, disconnected from the network.
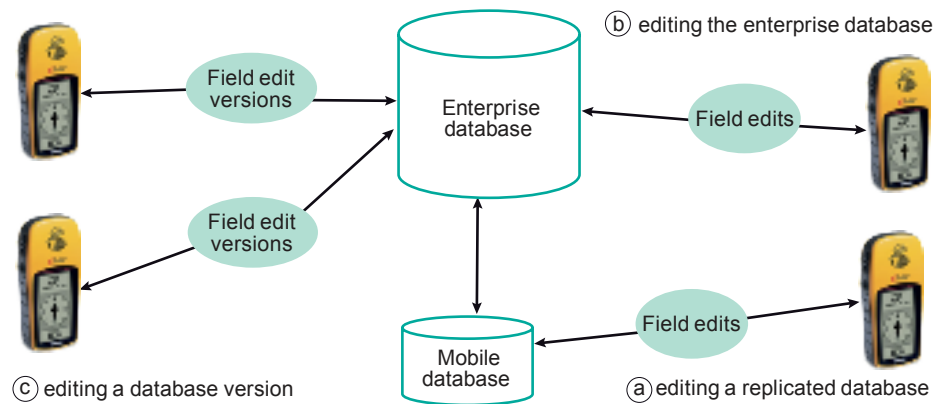
8.2.5 Digitizing and scanning of analogue maps

A traditional method of obtaining spatial data is through digitizing existing printed maps. This can be done using various techniques. Before adopting this approach, one must be aware that positional errors already on the map will further accumulate and that one must be willing to accept these errors.

There are two forms of digitizing: on-tablet and on-screen manual digitizing (Figure 8.30). In on-tablet digitizing, the original map is fitted on a special surface (the tablet), while in on-screen digitizing, a scanned image of the map (or some other image) is shown on the computer screen. In both of these forms, an operator follows the map's features (mostly lines) with a mouse device, thereby tracing the lines, and storing location coordinates relative to a number of previously defined control points.

The function of these points is to "lock" a coordinate system onto the digitized data: the control points on the map have known coordinates, so by digitizing them we tell the system implicitly where all other digitized locations are. At least three control points are needed, but preferably more should be digitized, to allow checking for any positional errors.

Figure 8.30
Manual digitizing techniques: (a) on-tablet digitizing; (b) on-screen digitizing.

Another set of techniques also works from a scanned image of the original map, but uses a GIS to find features in the image. These techniques are known as semi-automatic or automatic digitizing, depending on how much operator interaction is required. If vector data are to be distilled from this procedure, a process known as vectorization follows the scanning process. This procedure is less labour-intensive but can only be applied for relatively simple sources.

Scanning

A scanned image of the original map is needed for on-screen manual digitizing and semi-automatic/automatic digitizing. A range of scanners are available for obtaining a scanned image, starting from a small-format (A4) desktop scanner with resolutions of 200–800 dpi, through to high-end flatbed and drum scanners suitable for very accurate scanning of large-sized documents (A0) (Figure 8.31).
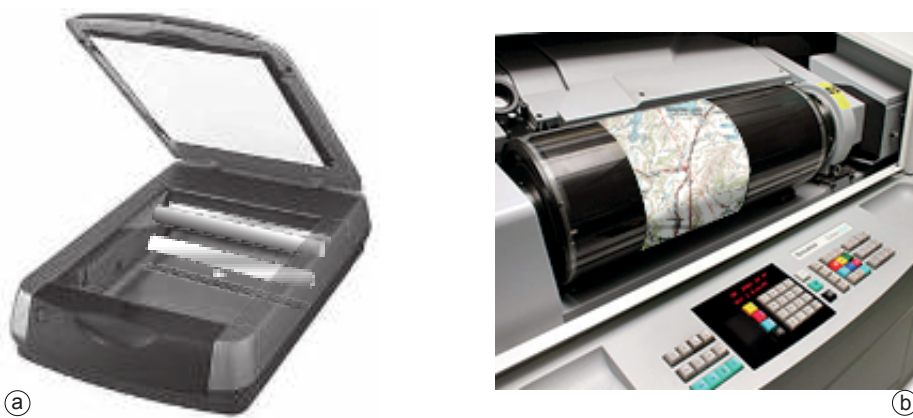
Figure 8.31
Main types of scanners: (a) a flatbed scanner. (b) a drum scanner.

A scanner illuminates a document and measures the intensity of the reflected light with a CCD array. The result is an image represented as a matrix of pixels, each of which holds an intensity value. Office scanners have a fixed maximum resolution, expressed as the highest number of pixels they can identify per inch; the unit is dots per inch (dpi). For manual on-screen digitizing of a printed map, a resolution of 200–300 dpi is usually sufficient, depending on the thickness of the thinnest lines. For man-

ual on-screen digitizing of aerial photographs, higher resolutions are recommended—typically, at least 800 dpi.

Semi-automatic/automatic digitizing requires a resolution that results in scanned lines of several pixels wide to enable the computer to trace the centre of the lines and thus avoid displacements. For printed maps, a resolution of 300–600 dpi is usually sufficient. Automatic or semi-automatic tracing from aerial photographs can only be done in a limited number of cases. Usually the information from aerial photos is obtained through visual interpretation.

After scanning, the resulting image can be improved by various image processing techniques. It is important to understand that scanning does not result in a structured data set of classified and coded objects. Additional work is required to recognize features and to associate categories and other thematic attributes with them.

Vectorization

The process of distilling points, lines and polygons from a scanned image is called vectorization. As scanned lines may be several pixels wide, they are often first thinned to retain only the centreline. The remaining centreline pixels are converted to series of (x, y) coordinate pairs, defining a polyline. Subsequently, features are formed and attributes are attached to them. This process may be entirely automated or performed semi-automatically, with the assistance of an operator. Pattern recognition methods—like Optical Character Recognition (OCR) for text—can be used for the automatic detection of graphic symbols and text.

Vectorization causes errors such as small spikes along lines, rounded corners, errors in T- and X-junctions, displaced lines or jagged curves. These errors are corrected in an automatic or interactive post-processing phase. The phases of the vectorization process are illustrated in Figure 8.32.

Scanned image   Noise

Vectorized data

After post-processing   Cleared
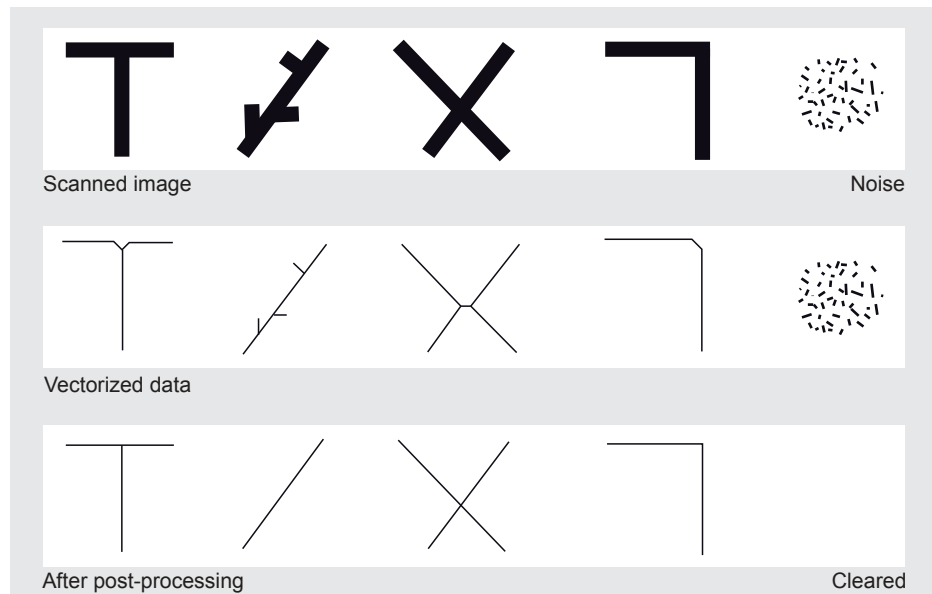
Figure 8.32
The phases of the vectorization process and various sorts of minor errors resulting from it. These are repaired in a post-processing phase.

Selecting a digitizing technique

The choice of digitizing technique depends on the quality, complexity and contents of the input document. Complex images are better manually digitized; simple images

are better automatically digitized. Images that are full of detail and symbols—such as topographic maps and aerial photographs—are therefore better digitized manually. Images that show only one type of information (e.g. elevation contours) are better automatically digitized.

In practice, the optimal choice may be a combination of methods. For example, contour-line film separations can be automatically digitized and used to produce a DEM. Existing topographic maps must be digitized manually, but new, geometrically corrected aerial photographs, with vector data from the topographic maps displayed directly over it, can be used for updating existing data files by means of manual on-screen digitizing.

8.2.6 Obtaining spatial data elsewhere

Over the past two decades, spatial data have been collected in digital form at an increasing rate and stored in various databases by the individual producers for their own use and for commercial purposes. More and more of these data are being shared among GIS users. There are several reasons for this. Some data are freely available, yet other data are only available commercially, as is the case for most satellite imagery. High quality data remain both costly and time consuming to collect and verify, as well as the fact that more and more GIS applications are looking at not just local, but national or even global, processes. As we will see in Section 8.4, new technologies have played a key role in the increasing availability of geospatial data. As a result of this increasing availability, we have to be more and more careful that the data we have acquired are of sufficient quality to be used in analyses and decision-making.

There are several related initiatives in the world to supply base data sets at national, regional and global levels, as well as those aiming to harmonize data models and definitions of existing data sets. Global initiatives include, for example, the Global Map, the USGS Global GIS database and the Second Administrative Level Boundaries (SALB) project. SALB, for instance, is a UN project aiming at improving the availability of information about administrative boundaries in developing countries.

Data formats and standards An important problem in any environment involved in digital data exchange is that of data formats and data standards. Different formats have been implemented by various GIS vendors, and different standards came about under different standardization committees. The phrase "data standard" refers to an agreed way, in terms of content, type and format, of representing data in a system. The good news about both formats and standards is that there are many to choose from; the bad news is that this can lead to a range of conversion problems. Several meta-data standards for digital spatial data exist, including those of the International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC).

8.3 Data preparation

Spatial data preparation aims to make acquired spatial data fit for use. Images may require enhancements and corrections of the classification scheme of the data. Vector data also may require editing, such as the trimming of line overshoots at intersections, deleting duplicate lines, closing gaps in lines, and generating polygons. Data may require conversion to either vector or raster formats to match other data sets that will be used in analyses. Additionally, the data preparation process includes associating attribute data with the spatial features through either manual input or reading digital attribute files into the GIS/DBMS.

The intended use of the acquired spatial data may require a less-detailed subset of the original data set, as only some of the features are relevant for subsequent analysis or subsequent map production. In these cases, data and/or cartographic generalization can be performed on the original data set.

This entire section treats a range of procedures for data checking, cleaning up, and integration to prepare vector data for analysis. Issues related to the preparation process of remote sensing data have already been discussed in Chapter 5.

8.3.1 Data checks and repairs

Acquired data sets must be checked for quality in terms of the accuracy, consistency and completeness. Often, errors can be identified automatically, after which manual editing methods can be used to correct the errors. Alternatively, some software may identify and automatically correct certain types of errors. The geometric, topological, and attribute components of spatial data are discussed in the following subsections.

automatic and manual checking

"Clean-up" operations are often performed in a standard sequence. For example, crossing lines are split before dangling lines are erased, and nodes are created at intersections before polygons are generated; see Figure 8.33.

	Before cleanup	After cleanup	Description
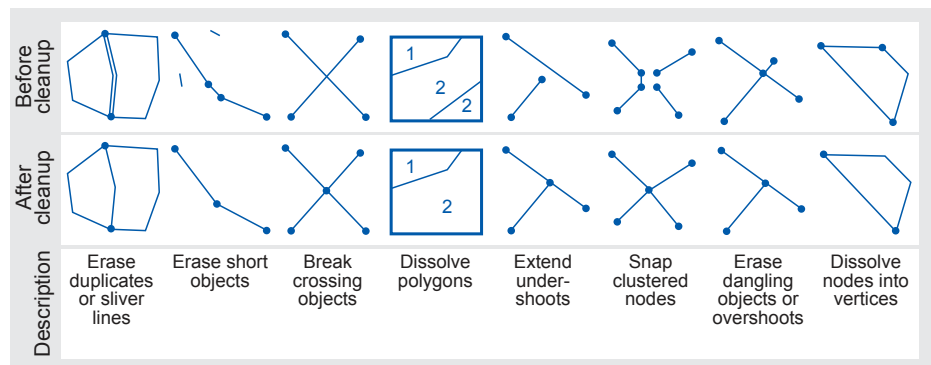
Figure 8.33
Clean-up operations for vector data.

| Erase duplicates or sliver lines | Erase short objects | Break crossing objects | Dissolve polygons | Extend under-shoots | Snap clustered nodes | Erase dangling objects or overshoots | Dissolve nodes into vertices |

With polygon data, one usually starts with many polylines (in an unwieldy format known as spaghetti data) that are combined and cleaned in the first step (Figure 8.34a–b). This results in fewer polylines with nodes being created at intersections. Then, polygons can be identified (Figure 8.34c). Sometimes, polylines that should connect to form closed boundaries do not, and must, therefore, be connected either manually or automatically. In a final step, the elementary topology of the polygons can be derived (Figure 8.34d).

Associating attributes   Attributes may be automatically associated with features that have unique identifiers (Figure 8.35). In the case of vector data, attributes are
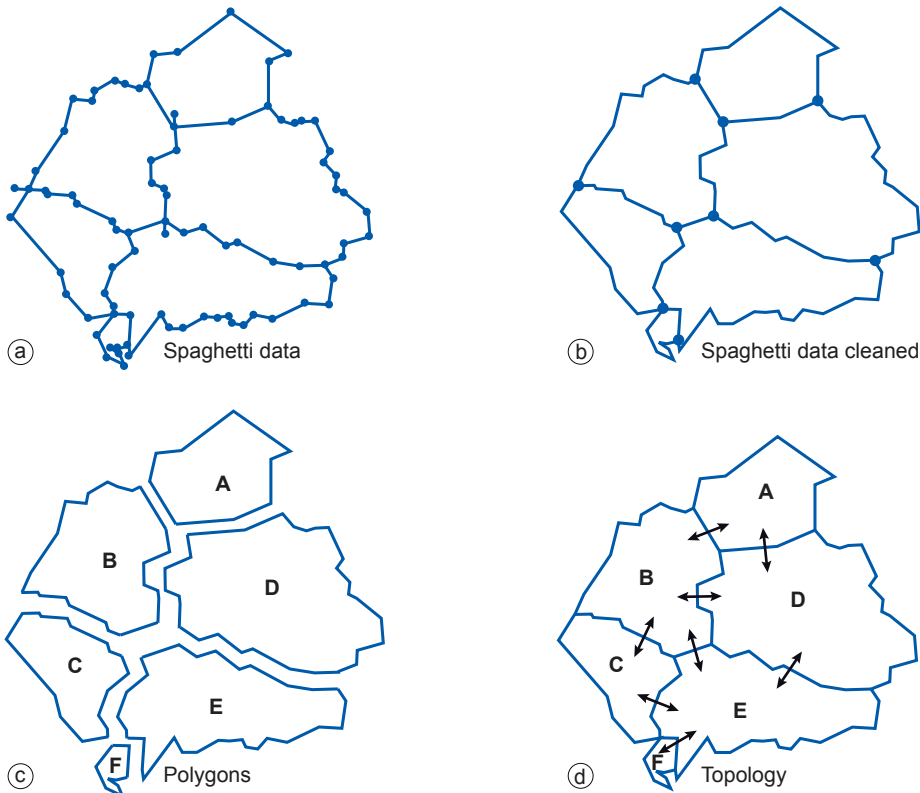
Figure 8.34
Successive clean-up operations for vector data, turning spaghetti data into topological structure.

assigned directly to the features, while in a raster the attributes are assigned to all cells that represent a feature.
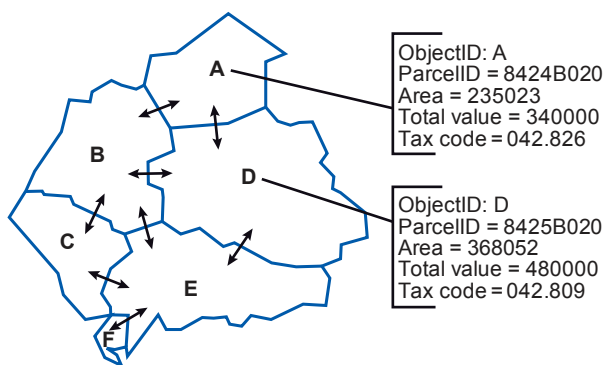
Figure 8.35
Attributes are associated with features that have unique identifiers.

It follows that, depending on the data type, assessment of attribute accuracy may range from a simple check on the labelling of features—for example, is a road classi-fied as a metalled road actually surfaced or not?—to complex statistical procedures for assessing the accuracy of numerical data, such as the percentage of pollutants present in a soil.

Rasterization or vectorization   Vectorization produces a vector data set from a raster. In some sense, we have looked at this already: namely in the production of

a vector set from a scanned image. Another form of vectorization is used when we want to identify features or patterns in remotely sensed images. The keywords here are feature extraction and pattern recognition, which are dealt with in Chapter 6.

If much or all of the subsequent spatial data analysis is to be carried out on raster data, one may want to convert vector data sets to raster data. This process, known as rasterization, involves assigning point, line and polygon attribute values to raster cells that overlap with the respective point, line or polygon. To avoid information loss, the raster resolution should be carefully chosen on the basis of the geometric resolution. A cell size that is too large may result in cells that cover parts of multiple vector features, and then ambiguity arises as to what value to assign to the cell. If, on the other hand, the cell size is too small, the file size of the raster may increase significantly.

Rasterization itself could be seen as a "backwards step": firstly, raster boundaries are only an approximation of the objects' original boundary. Secondly, the original "objects" can no longer be treated as such, as they have lost their topological properties. Rasterization is often done because it facilitates easier combination with other data sources that are also in raster formats, and/or because there are several analytical techniques that are easier to perform on raster data (see Chapter 9). An alternative to rasterization is to not perform it during the data preparation phase, but to use GIS rasterization functions "on the fly", i.e. when the computations call for it. This allows the vector data to be kept and raster data to be generated from them when needed. Obviously, the issue of performance trade-offs must be looked into.

Topology generation   We have already discussed the derivation of elementary polygon topology starting from uncleaned polylines. However, more topological relations may sometimes be needed, as for instance in networks where questions of line connectivity, flow direction and which lines have overpasses and underpasses may need to be addressed. For polygons, questions that may arise involve polygon inclusion: is a polygon inside another one, or is the outer polygon simply around the inner polygon?

In addition to supporting a variety of analytical operations, topology can aid in data editing and in ensuring data quality. For example, adjacent polygons such as parcels have shared edges; they do not overlap, nor do they have gaps. Typically, topology rules are first defined (e.g. "there should be no gaps or overlap between polygons"), after which validation of the rules takes place. The topology errors can be identified automatically, to be followed by manual editing methods to correct the errors.

An alternative to storing topology together with features in the spatial database is to create topology on the fly, i.e. when the computations call for it. The created topology is temporary, only lasting for the duration of the editing session or analysis operation.

8.3.2 Combining data from multiple sources

A GIS project usually involves multiple data sets, so the next step addresses the issue of how these multiple sets relate to each other. The data sets may be of the same area but differ in accuracy, or they may be of adjacent areas, having been merged into a single data set, or the data sets may be of the same or adjacent areas but are referenced in different coordinate systems. Each of these situations is discussed in the following subsections.

Differences in accuracy   Issues relating to positional error, attribute accuracy and temporal accuracy are clearly relevant in any combination of data sets, which may themselves have varying levels of accuracy.

Images come at a certain resolution, and printed maps at a certain scale. This typically

results in differences of resolution of acquired data sets, all the more since map features are sometimes intentionally displaced or in another way generalized to improve readability of the map. For instance, the course of a river will only be approximated roughly on a small-scale map, and a village on its northern bank should be depicted north of the river, even if this means it has to be displaced on the map a little bit. The small scale causes an accuracy error. If we want to combine a digitized version of that map with a digitized version of a large-scale map, we must be aware that features may not be where they seem to be. Analogous examples can be given for images of different resolutions.
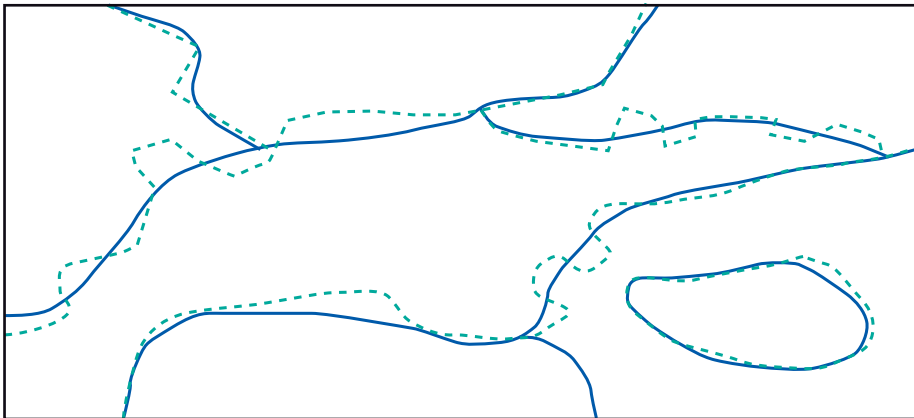
Figure 8.36
The integration of two vector data sets representing the same phenomenon may lead to sliver polygons.

In Figure 8.36, the polygons of two digitized maps at different scales are overlaid. Owing to scale differences in the sources, the resulting polygons do not perfectly coincide, and polygon boundaries cross each other. This causes small, artefact polygons in the overlay that are known as sliver polygons. If the map scales differ significantly, the polygon boundaries of the large-scale map should probably take priority, but when the differences are slight, we need interactive techniques to resolve any issues.

sliver polygons

There can be good reasons for having data sets at different scales. A good example is found in mapping organizations. European organizations maintain a single source database that contains the base data. This database is essentially scale-less and contains all data required for even the largest scale map to be produced. For each map scale that the mapping organization produces, they derive a separate database from the foundation data. Such a derived database may be called a cartographic database since the data stored are elements to be printed on a map, including, for instance, data on where to place name tags and what colour to give them. This may mean the organization has one database for the larger scale ranges (1:5000–1:10,000) and other databases for the smaller scale ranges; they maintain a multi-scale data environment.

More recent research has addressed the development of one database incorporating both larger and smaller scale ranges. Here we identify two main approaches: one approach to realize this is to store multiple representations of the same object in a multiple representation database. The database must keep track of links between different representations for the same object and must also provide support for decisions as to which representations to use in which situation. Another approach is to maintain one database for the larger scale ranges and derive representations for the smaller scale ranges on the fly. That means the data have to be generalized in real time. A combination of both approaches is to store multiple object representations for time-consuming generalization processes, which sometimes cannot be done fully automatically, and derive representations for the smaller scales in real time on the fly for rapid general-

ization processes.

Merging data sets of adjacent areas   When individual data sets have been prepared as just described, they sometimes have to be integrated into a single, "seamless" data set, whilst ensuring that the appearance of the integrated geometry is as homogeneous as possible. Edge matching is the process of joining two or more map sheets, for instance, after they have been separately digitized.

<div style="float:left">edge matching</div>

Merging adjacent data sets can be a major problem. Some GIS functions, such as line smoothing and data clean-up (removing duplicate lines) may have to be performed. Figure 8.37 illustrates a typical situation. Some GISs have merge or edge-matching functions to solve the problems arising from merging adjacent data. At map-sheet edges, feature representations have to be matched in order for them to be combined.
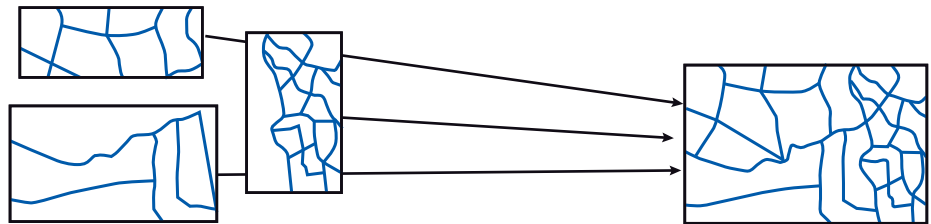
Figure 8.37
Multiple adjacent data sets, after cleaning, can be matched and merged into a single data set.

Coordinates of the objects along shared borders are adjusted to match those in the neighbouring data sets. Mismatches may still occur, so a visual check and interactive editing is likely to be required.

Differences in coordinate systems   It may be the case that data layers that are to be combined or merged in some way are referenced in different coordinate systems, or are based upon different datums. As a result, data may need a coordinate transformation (Figure 3.22), or both a coordinate transformation and datum transformation (Figure 3.23). It may also be the case that data have been digitized from an existing map or data layer (Subsection 8.2.6). In this case, geometric transformations help to transform device coordinates (coordinates from digitizing tablets or screen coordinates) into world coordinates (geographic coordinates, metres, etc.).
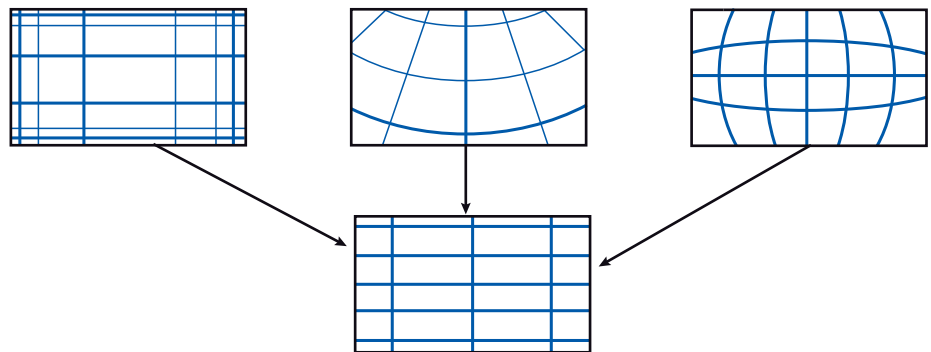
Figure 8.38
The integration of data sets into one common coordinate system.

Other data preparation functions   A range of other data preparation functions exist that support conversion or adjustment of the acquired data to format requirements

that have been defined for data storage purposes. These include format transformation functions; functions to bring data sets in line with a common data-content model or a particular database schema; functions for removal of redundant vertices; data clipping operations; and so on. These functions and others discussed earlier are often performed in a standard sequence.

A good illustration of applying data preparation functions in a standard order is the SALB project. In this project, editing procedures were developed to clean the delimitation of the administrative units—to make them fit the international border standard—and prepare the attributes. These procedures are standardized to ensure the comparability between countries in terms of quality. The preparation functions for these procedures include changing formats and projections; adding missing delimitation of administrative boundaries; clipping the administrative boundaries to the international border; uniting multiple polygons; deleting redundant vertices; cleaning and checking attributes; building topology for the units; correcting gaps and overlaps between the units; and calculating attribute values such as areas and lengths.

8.4 Data management and processing systems

The ability to manage and process spatial data is a critical component for any functioning GIS. Simply put, data processing systems refer to hardware and software components that are able to process, store and transfer data. This section discusses the components of systems that facilitate the handling of spatial data and processing of geoinformation. To provide some context for the discussion, the section begins with a brief discussion of computer hardware and software trends over the past three decades.

In the subsections that follow, we then discuss database management systems (DBMSs) and illustrate some principles and methods of data extraction from a database. The final subsection of this section (Subsection 8.4.3) looks at the merging of GISs and DBMSs, and the emergence of spatial databases in recent years. It notes their key advantages, and briefly illustrates the use of a spatial database for data storage and processing. Before we deal with database aspects in detail, however, it is good to review the spatial data handling process, since it puts constraints on how we intend to organize our data and what kind of questions the system should be able to answer.

8.4.1 Stages of spatial data handling

Spatial data capture and preparation

Functions for capturing data are closely related to the disciplines of surveying engineering, photogrammetry, and remote sensing and the processes of digitizing. Remote sensing, in particular, is the field that provides photographs and images as the raw, base data from which spatial data sets are derived. Surveys of a study area often need to be conducted to collect data that cannot be obtained with remote sensing techniques, or to validate the data thus obtained. Traditional techniques for obtaining spatial data, typically from paper sources, included manual digitizing and scanning. In recent years there has been a significant increase in the availability and sharing of digital—geospatial—data. Various media and computer networks play an important role in the dissemination of these data, particularly the Internet.

The data, even though it has been obtained in a digital format, may still not be quite ready for use in the system. This may be because the format applied in the capturing process was not quite the format required for storage and further use; some type of data conversion is then required. In part, this problem may also arise when the captured data represent only raw, base data, from which the data objects of real interest to the system still need to be constructed. For example, semi-automatic digitizing may produce line segments, while the application requires non-overlapping polygons. A build and verification phase would then be needed to obtain these from the captured lines. Issues related to data acquisition and preparation are discussed in greater detail in the Sections 8.2 and 8.3.

Spatial data storage and maintenance

The way that data are stored plays a central role in their processing and, eventually, our understanding of them. In most available systems, spatial data are organized in layers by theme and/or scale. Examples are layers of thematic categories, such as land use, topography and administrative subdivisions, each according to their mapping scale. An important underlying principle is that a representation of the real world has to be designed such that it reflects phenomena and their relationships as naturally as possible. In a GIS, features are represented together with their attributes—geometric and non-geometric—and relationships. The geometry of features is represented with primitives of the respective dimension: a windmill probably as a point; an agricultural

field as a polygon. The primitives follow either the vector or the raster approach.

As discussed in Section 8.1, vector data types describe an object through its boundary, thus dividing the space into parts that are occupied by the respective objects. The raster approach subdivides space into (regular) cells, mostly as a square tessellation of two or three dimensions. These cells are called pixels in 2D and voxels in 3D. The data indicate for every cell which real-world feature is covered, provided the cell represents a discrete field. In the case of a continuous field, the cell holds a representative value for that field. Table 8.2 lists advantages and disadvantages of raster and vector representations.

Raster representation	Vector representation
advantages	
• simple data structure	• efficient representation of topology
• simple implementation of overlays	• adapts well to scale changes
• efficient for image processing	• allows representing networks
	• allows easy association with attribute data
disadvantages	
• less compact data structure	• complex data structure
• difficulties in representing topology	• overlay more difficult to implement
• cell boundaries independent of feature boundaries	• inefficient for image processing
	• more update-intensive

Table 8.1
Raster and vector
representations compared.

The storage of a raster is, in principle, straightforward. It is stored in a file as a long list of values, one for each cell, preceded by a small list of extra data (the "file header"), which specifies how to interpret the long list. The order of the cell values in the list can, but need not necessarily, be left to right, top to bottom. This simple encoding scheme is known as row ordering. The header of the raster will typically specify how many rows and columns the raster has, which encoding scheme was used, and what sort of values are stored for each cell.

header file

Raster files can be large. For efficiency reasons, it is wise to organize the long list of cell values in such a way that spatially nearby-cells are also near to each other in the list. This is why other encoding schemes have been devised. The reader is referred to [65] for a more detailed discussion.

Low-level storage structures for vector data are much more complicated, and a discussion of this topic is beyond the scope of this textbook. The best intuitive understanding can be obtained from Figure 8.10, which illustrates a boundary model for polygon objects. Similar structures are in use for line objects. For further, advanced, reading see [102]. GIS packages support both spatial and attribute data, i.e. they accommodate spatial data storage using a vector approach and attribute data using tables. Historically, however, database management systems (DBMSs) have been based on the notion of tables for data storage.

GIS applications have been able to link to an external database to store attribute data and make use of its superior data management functions. Currently, all major GIS packages provide facilities to link with a DBMS and exchange attribute data with it. Spatial (vector) and attribute data are still sometimes stored in separate structures, although they can now be stored directly in a spatial database. More detail on these issues is provided in Subsection 8.4.2. Maintenance of data, spatial or otherwise, can best be defined as the combination of activities needed to keep the data set up to date

and as supportive as possible for the user community. It deals with obtaining new data and entering them into the system, as well as possibly replacing outdated data. The purpose is to have an up-to-date, stored data set available. After a major flood, for instance, we may have to update road-network data to reflect that roads have been washed away or have become otherwise impassable.

The need for updating spatial data originates from the requirements posed by the users, as well as the fact that many aspects of the real world change continuously. Data updates can take different forms. It may be that a completely new survey has been carried out, from which an entirely new data set will be derived, to replace the current set. This is typically the case if the spatial data originate from remote sensing, for example, from a new vegetation-cover set or from a new digital elevation model. Furthermore, local ground surveys may reveal local changes, such as new constructions or changes in land use or ownership. In such cases, local changes to a large spatial data set are typically required. Such local changes should take into account matters of data consistency, i.e. they should leave other spatial data within the same layer intact and correct.

Spatial query and analysis

The most characteristic parts of a GIS are its functions for spatial analysis, i.e. operators that use spatial data to derive new geoinformation. Spatial queries and process models play an important role in this functionality. One of the key uses of GISs has been to support spatial decision-making. Spatial decision-support systems (SDSSs) are a category of information systems composed of a database, GIS software, models, and a "knowledge engine" that allows users to deal specifically with location-related problems.

GIS functions are used for maintenance of the data and for analysing the data in order to infer spatial information. Analysis of spatial data can be defined as computing new information to provide new insights from existing spatial data. Consider an example from the domain of road construction. In mountainous areas, this is a complex engineering task with many cost factors, including the number of tunnels and bridges to be constructed, the total length of the tarmac, and the volume of rock and soil to be moved. GISs can help to compute such costs on the basis of an up-to-date digital elevation model and a soil map. Maintenance and analysis of attribute data is discussed further in Subsection 8.4.2. The exact nature of the analysis will depend on the application requirements, but computations and analytical functions can operate on both spatial and non-spatial data; Chapter 9 discusses these issues in more detail. For now, we will focus on the last stage of Figure 8.39, the presentation of spatial data.

Spatial data presentation

The presentation of spatial data, whether in print or on-screen, on maps or in tabular displays, or as "raw" data, is closely related to the discipline of cartography. The presentation may either be an end-product, for example a printed atlas, or an intermediate product, such as spatial data made available through the Internet. Table 8.2 lists several methods and devices used for the presentation of spatial data.

Cartography, information visualization and scientific visualization make use of these methods and devices in their products. Section 10.1 is devoted to visualization techniques for spatial data.

8.4.2 Database management systems

database

A database is a large, computerized collection of structured data. In the non-spatial domain, databases have been in use since the 1960s for various purposes, such as bank

Method	Devices
Hardcopy	• printer • plotter (pen plotter, ink-jet printer, thermal transfer printer, electrostatic plotter) • film writer
Soft copy	• computer screen
Output of digital data sets	• magnetic tape • CD-ROM or DVD • the Internet

Table 8.2
Spatial data presentation.

account administration, stock monitoring, salary administration, sales and purchasing administration and flight reservation systems. These applications have in common that the amount of data is quite large, but the data themselves have a simple and regular structure. Designing a database is not an easy task. First, one has to consider carefully what the purpose of the database is and who its users will be. Second, one needs to identify the available data sources and define the format in which the data will be organized within the database. This format is usually called the database structure. Only when all this is in place can data be entered into the database. Data must be kept up to date and it is, therefore, wise to set up the processes for doing this and to make someone responsible for regular maintenance. Documentation of the database design and set up is crucial for an extended database life (proprietary databases tend to outlive the professional careers of their original designers).

A database management system (DBMS) is a software package that allows the user to set up, use and maintain a database. Just as a GIS allows the set up of a GIS application, a DBMS offers generic functionality for database organization and data handling. In the next subsection we take a closer look at what type of functions are offered by DBMSs. Many standard PCs are equipped with a DBMS called Microsoft Access. This package offers a useful set of functions and the capacity to store terabytes of information.

DBMS

Reasons for using a DBMS

There are various reasons why one would want to use a DBMS for data storage and processing:

- A DBMS supports the storage and manipulation of very large data sets. Some data sets are so big that storing them in text files or spreadsheet files becomes too awkward for practical use. The result may be that finding simple facts takes minutes, and performing simple calculations perhaps even hours. A DBMS is specifically designed for these purposes.

- A DBMS can be instructed to guard data correctness. For instance, an important aspect of data correctness is data-entry checking: ensuring that the data that are entered into the database do not contain obvious errors. For instance, since we know in what study area we are working, we also know the range of possible geographic coordinates, so we can ensure the DBMS checks them upon entry. This is a simple example of the type of rules, generally known as integrity constraints, that can be defined in, and automatically checked by, a DBMS. More complex integrity constraints are certainly possible; their definition is is an aspect of the database design.

- A DBMS supports the concurrent use of the same data set by many users. Large data sets are often built up over time. As a result, substantial investments are required to create and maintain them, and probably many people are involved in the data collection, maintenance and processing. Such data sets are often considered to have high strategic value by their owner(s) and many people may want to use them within an organization. Moreover, different users may have different views about the data. As a consequence, users will be under the impression that they are operating on their personal database and not on one shared by many people. They may all be using the database at the same time without affecting each other's activities. This DBMS function is referred to as concurrency control.

- A DBMS provides users with a high-level, declarative query language, with as its most important use the formulation of queries.

- A DBMS supports the use of a data model, which is a language with which one can define a database structure and manipulate the data stored in it. The most prominent data model is the relational data model; this is discussed in full in Subsection 8.4.3. Its primitives are tuples (also known as records, or rows) with attribute values, and relations, which are sets of similarly formed tuples.

- A DBMS includes data backup and recovery functions, to ensure data availability at all times. As potentially many users rely on the availability of the data, the data must be safeguarded against possible calamities. Regular backups of the data set and automatic recovery schemes provide insurance against loss of data.

- A DBMS allows the control of data redundancy. A well-designed database takes care of storing single facts only once. Storing a fact several times—a phenomenon known as data redundancy—can lead to situations in which stored facts may contradict each other, causing reduced usefulness of the data. Redundancy is, however, not necessarily always problematic, as long as we specify where it occurs so that it can be controlled.

Alternatives for data management

The decision whether or not to use a DBMS will depend, among other things, on how much data there are or will be, what type of use will be made of it, and how many users might be involved. On the small-scale side of the spectrum—when the data set is small, its use is relatively simple, and there is just one user—we might use simple text files and a word processor. Think of a personal address book as an example or a small set of simple field observations. Text files offer no support for data analysis, except perhaps in alphabetical sorting. If our data set is still small and numeric in nature, and we have a single type of use in mind, a spreadsheet program might suffice. This might also be the case if we have a number of field observations with measurements that we want to prepare for statistical analysis.

If, however, we carry out region- or nation-wide censuses, with many observation stations and/or field observers and all sorts of different measurements, one quickly needs a database to keep track of all the data. Spreadsheet programs are generally not suitable for this, however, as they do not accommodate concurrent use of data sets well, although they do support some data analysis, especially when it comes to calculations for a single table, such as averages, sums, minimum and maximum values.

All such computations are usually restricted to a single table of data. When one wants to relate the values in the table with values of another nature in some other table, skilful expertise and significant amounts of time may be required to achieve this.

PrivatePerson	TaxId	Surname	BirthDate
	101-367	Garcia	10/05/1952
	134-788	Chen	26/01/1964
	101-490	Fakolo	14/09/1931

Parcel	PId	Location	AreaSize
	3421	2001	435
	8871	1462	550
	2109	2323	1040
	1515	2003	245

TitleDeed	Plot	Owner	DeedDate
	2109	101-367	18/12/1996
	8871	101-490	10/01/1984
	1515	134-788	01/09/1991
	3421	101-367	25/09/1996

Figure 8.39
An example of a small database consisting of three relations (tables), all with three attributes, and three, four and four tuples, respectively. PrivatePerson / Parcel / TitleDeed are the names of the three tables. Surname is an attribute of the PrivatePerson table; the Surname attribute value for person with TaxId '101-367' is 'Garcia'.

Relational data models

For relational data models, the structures used to define the database are attributes, tuples and relations. Computer programs either perform data extraction from the database without altering it, in which case they are termed queries, or they change the database contents, in which case we speak of updates or transactions. The technical terms related to database technology are defined below. An extremely small database selected from a cadastral setting is illustrated in Figure 8.39. This database consists of three tables, one for storing people's details, one for storing land-parcel details and a third for storing details concerning title deeds. Various sources of information are kept in the database such as a taxation identifier (TaxId) for people, a parcel identifier (PId) for land parcels, and the date of a title deed (DeedDate).

relational data model

PrivatePerson	(TaxId : string, Surname : string, Birthdate : date)
Parcel	(Pid : number, Location : polygon, AreaSize : number)
TitleDeed	(Plot : number, Owner : string, DeedDate : date)

Table 8.3
The relation schemas for the three tables of the database in Figure 8.39.

Relations, tuples and attributes   In relational data models, a database is viewed as a collection of relations, also commonly referred to as tables.

A data model is a language that allows the definition of:

relation

- the structures that will be used to store the base data;

- the integrity constraints that the stored data have to obey at all moments in time;

- the computer programs used to manipulate the data.

A table or relation is itself a collection of tuples (or records). In fact, each table is a collection of tuples that are similarly shaped. By this, we mean that a tuple has a fixed number of named fields (also known as attributes). All tuples in the same relation have the same named fields. In a diagram, such as in Figure 8.39, relations can be displayed as data in tabular form, as the relations provided in the figure demonstrate. The PrivatePerson table has three tuples; the Surname attribute value for the first tuple shown is "Garcia."

tuple

attribute domain

The phrase "that are similarly shaped" takes this a bit further. It requires that all values for the same attribute come from a single domain of values. An attribute's domain is a (possibly infinite) set of atomic values such as, for example, the set of integer number values or the set of real number values. In our cadastral database example, the domain of the Surname attribute, for instance, is a string, so any surname is represented as a sequence of text characters, i.e. as a string. The availability of other domains depends on the DBMS, but usually integer (the whole numbers), real (all numbers), date, yes/no and a few more are included. When a relation is created, we need to indicate what type of tuples it will store. This means that we must:

1. provide a name for the relation;

2. indicate which attributes it will have;

3. set the domain of each attribute.

relation schema

attribute

A relation definition obtained in this way is known as the relation schema of that relation. The definition of a relation schema is an important part of a database. An attribute is a named field of a tuple, with which each tuple associates a value, the tuple's attribute value. Our example database has three relation schemas; one of which is TitleDeed. The relation schemas together make up the database schema. The relation schemas for the database of Figure 8.39 are given in Table 8.3. Underlined attributes (and their domains) indicate the primary key of the relation, which will be defined and discussed below.

Relation schemas are stable and will rarely change over time. This is not true of the tuples stored in tables: typically, they are often changing, either because new tuples are added or others are removed, or still others will undergo changes in their attribute values. The set of tuples in a relation at some point in time is called the relation instance at that moment. This tuple set is always finite: you can count how many tuples there are. Figure 8.39 gives us a single database instance, i.e. one relation instance for each relation. One of the relation instances has three tuples, two of them have four. Any relation instance always contains only tuples that comply with the relation schema of the relation.

Finding tuples and building links between them   We have already indicated that database systems are particularly good at storing large quantities of data. (Our example database is not even small, it is tiny!) The DBMS must support rapid searching among many tuples. This is why relational data models use the notion of a key. In other words, if we have a value for each of the key attributes we are guaranteed to find no more than one tuple in the table with that combination of values; it remains possible that there is no tuple for the given combination. In our example database, the set TaxId, Surname is a key of the relation PrivatePerson: if we know both a TaxId and a Surname value, we will find at most one tuple with that combination of values.

Every relation has a key, though possibly it is the combination of all attributes. Such a large key is, however, not handy because we must provide a value for each of its attributes when we search for tuples. Clearly, we want a key to have as few as possible attributes: the fewer, the better.

A key of a relation comprises one or more attributes. A value for these attributes uniquely identifies a tuple.

key

If a key has just one attribute, it obviously cannot have less attributes. Some keys have two attributes; an example is the key Plot, Owner of relation TitleDeed. We need
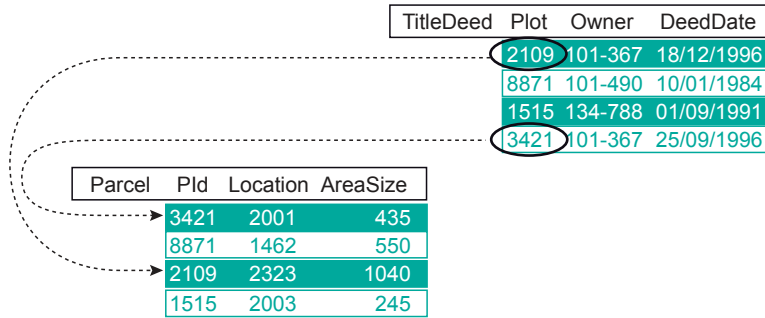
both attributes because there can be many title deeds for a single plot (in the case of plots that are sold often), but also many title deeds for a single person (say, in the case of wealthy persons). When we provide a value for a key, we can look up the corresponding tuple in the table (if such a tuple exists). A tuple can refer to another tuple by storing that other tuple's key value. For instance, a TitleDeed tuple refers to a Parcel tuple by including that tuple's key value. The TitleDeed table has a special attribute Plot for storing such values. The Plot attribute is called a foreign key because it refers to the primary key (Pid) of another relation (Parcel). This is illustrated in Figure 8.41.

foreign key

Two tuples of the same relation instance can have identical foreign key values: for instance, two TitleDeed tuples may refer to the same Parcel tuple. A foreign key is, therefore, not a key of the relation in which it appears, despite its name! A foreign key must have as many attributes as the primary key that it refers to.
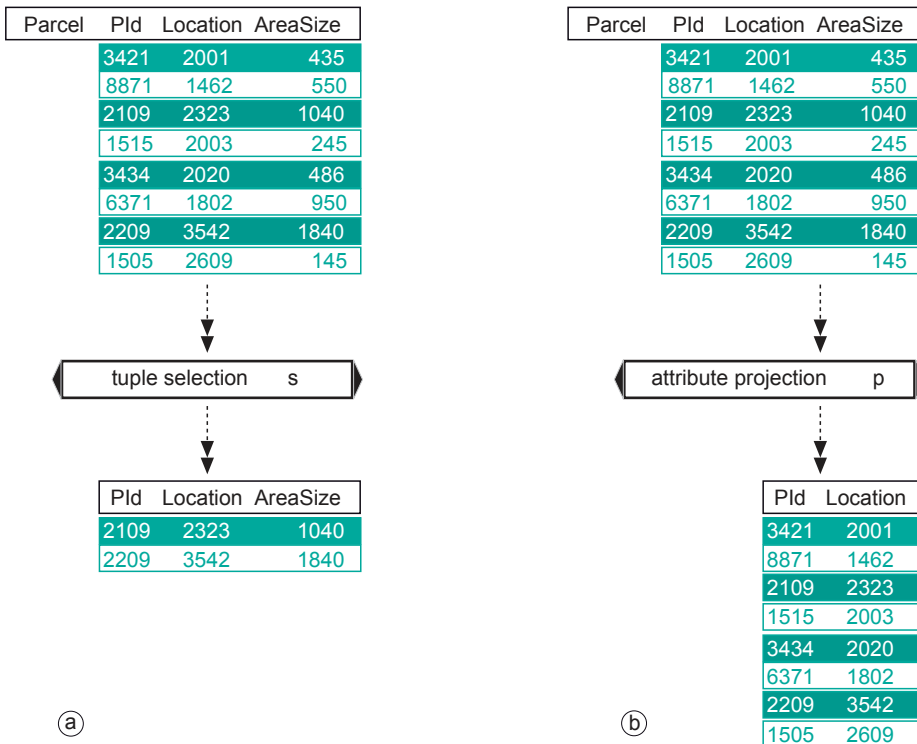
Querying a relational database

We will now look at the three most elementary query operators. These are quite powerful because they can be combined to define queries of higher complexity. The three query operators have some common traits. First, all of them require input and produce output, and both input and output are relations! This guarantees that the output of one query (a relation) can be the input of another query, which makes it possible to build more and more complex queries—if we want to.

The first query operator is called tuple selection. Tuple selection works like a filter: it allows tuples that meet the selection condition to pass and disallows tuples that do not meet the condition; see Figure 8.41a. The operator is given some input relation, as well as a selection condition about tuples in the input relation. A selection condition is a truth statement about a tuple's attribute values, such as AreaSize > 1000. For some tuples in Parcel, this statement will be true and for others it will be false. Tuple selection on the Parcel relation with this condition will result in a set of Parcel tuples for which the condition is true.

A second operator, called attribute projection, is also illustrated in Figure 8.41. Besides an input relation, this operator requires a list of attributes, all of which should be attributes of the schema of the input relation. The output relation of this operator has as its schema only the list of attributes given, so we say that the operator projects onto these attributes. Contrary to the first operator, which produces fewer tuples, this operator produces fewer attributes compared to the input relation.

The most common operator for defining queries in a relational database is the language SQL, which stands for Structured Query Language. The two queries of Figure 8.41 would be written in SQL as follows:

SQL

```
SELECT    *                    SELECT    Pid, Location
FROM      Parcel               FROM      Parcel
WHERE     AreaSize > 1000
```

(a) tuple selection from the Parcel relation, using the condition AreaSize > 1000. The * indicates that we want to extract all attributes of the input relation.

(b) attribute projection from the Parcel relation. The SELECT clause indicates that we only want to extract the two attributes Pid and Location. There is no WHERE-clause in this query.

Attribute projection works like a tuple formatter: it passes through all tuples of the input, and reshapes each of them in the same way.

Queries like the two above do not create stored tables in the database. This is why the result tables have no name: they are virtual tables. The result of a query is a table that is shown to the user who executed the query. Whenever the user closes her/his view on the query result, that result is lost. The SQL code for the query is, however, stored for future use. The user can re-execute the query again to obtain a view on the result once more.

SQL differs from the other two query languages in that it requires two input relations. The operator is called the join and is illustrated in Figure 8.42.

The output relation of this operator has as attributes those of the first and the second input relations. The number of attributes therefore increases. The output tuples are obtained by taking a tuple from the first input relation and "gluing" it to a tuple from the second input relation. The join operator uses a condition that expresses which tuples from the first relation are combined ('glued') with which tuples from the second. The example of Figure 8.42 combines TitleDeed tuples with Parcel tuples, but only

those for which the foreign key Plot matches with primary key PId.

The join operator takes two input relations and produces one output relation, gluing two tuples together (one from each input relation) to form a bigger tuple—provided they meet a specified condition.

The join query for our example is easily expressed in SQL as:

SELECT ?
FROM TitleDeed, Parcel
WHERE TitleDeed.Plot = Parcel.PId

The FROM clause identifies the two input relations; the WHERE clause states the join condition.

TitleDeed	Plot	Owner	DeedDate
	2109	101-367	18/12/1996
	8871	101-490	10/01/1984
	1515	134-788	01/09/1991
	3421	101-367	25/09/1996

Parcel	PId	Location	AreaSize
	3421	2001	435
	8871	1462	550
	2109	2323	1040
	1515	2003	245

join ⋈

Plot	Owner	DeedDate	PId	Location	AreaSize
2109	101-367	18/12/1996	2109	2323	1040
8871	101-490	10/01/1984	8871	1462	550
1515	134-788	01/09/1991	1515	2003	245
3421	101-367	25/09/1996	3421	2001	435

Figure 8.42
The essential binary query operator: join. The join condition for this example is TitleDeed.Plot = ParcelPid, which expresses a foreign key/key link between TitleDeed and Parcel. The result relation has $3 + 3 = 6$ attributes.

It is often not sufficient to use just one operator for extracting sensible information from a database. The strength of the above operators is hidden in the fact that they can be combined to produce more advanced and useful query definitions. A final example illustrates this. Take another look at the join of Figure 8.42. Suppose we really wanted to obtain combined TitleDeed/Parcel information, but only for parcels with a size over 1000, and we only wanted to see the owner identifier and deed date of such title deeds.

We can take the result of the join above and select the tuples that show a parcel size over 1000. The result of this tuple selection can then be taken as the input for an attribute selection that only leaves Owner and DeedDate. This is illustrated in Figure 8.43.

Finally, we may look at the SQL statement that would give us the query of Figure 8.43. It can be written as:

SELECT Owner
DeedDate FROM TitleDeed
Parcel WHERE TitleDeed.Plot = Parcel.PId AND AreaSize > 1000

8.4.3   GISs and spatial databases

Linking GISs and DBMSs

GIS software provides support for spatial data and thematic or attribute data. GISs have traditionally stored spatial data and attribute data separately. This required the GIS to provide a link between the spatial data (represented with rasters or vectors), and their non-spatial attribute data. The strength of GIS technology lies in its built-in "understanding" of geographic space and all functions that derive from this, for purposes such as storage, analysis and map production. GIS packages themselves can store tabular data, but they do not always provide a fully-fledged query language to operate on the tables.

TitleDeed	Plot	Owner	DeedDate
	2109	101-367	18/12/1996
	8871	101-490	10/01/1984
	1515	134-788	01/09/1991
	3421	101-367	25/09/1996

Parcel	PId	Location	AreaSize
	3421	2001	435
	8871	1462	550
	2109	2323	1040
	1515	2003	245

join   ⋈

Plot	Owner	DeedDate	PId	Location	AreaSize
2109	101-367	18/12/1996	2109	2323	1040
8871	101-490	10/01/1984	8871	1462	550
1515	134-788	01/09/1991	1515	2003	245
3421	101-367	25/09/1996	3421	2001	435

tuple selection   σ

Plot	Owner	DeedDate	PId	Location	AreaSize
2109	101-367	18/12/1996	2109	2323	1040

attribute projection   π

Owner	DeedDate
101-367	18/12/1996

Figure 8.43
A combined selection/projection/join query for selecting owners and deed dates for parcels with a size larger than 1000. The join is done first, then follows a tuple selection on the resulting tuples of the join, which is completed by an attribute projection.

DBMSs have a long tradition in handling attribute (i.e. administrative, non-spatial, tabular, thematic) data in a secure way for multiple users at the same time. Arguably, DBMSs offer much better table functionality, since they are specifically designed for this purpose. Many data in GIS applications are attribute data, so it made sense to use a DBMS for it. For this reason, many GIS applications have made use of external

DBMSs for data support. In this role, the DBMS serves as a centralized data repository for all users, while each user runs her/his own GIS software, which obtains its data from the DBMS. This means that the GIS has to link the spatial data represented by rasters or vectors and the attribute data stored in an external DBMS.

With raster representations, each raster cell stores a characteristic value. This value can be used to look up attribute data in an accompanying database table. For instance, the land use raster of Figure 8.44 indicates the land use class for each of its cells, while an accompanying table provides full descriptions for all classes, perhaps including some statistical information for each of the types. Note the similarity with the key/foreign key concept in relational databases.

With vector representations, our spatial objects—whether they are points, lines or polygons—are automatically given a unique identifier by the system. This identifier is usually just called the *object ID* or feature ID and is used to link the spatial object (as represented by vectors) with its attribute data in an attribute table. The principle applied here is similar to that in raster settings, but in this case each object has its own identifier. The ID in the vector system functions as a key, and any reference to an ID value in the attribute database is a foreign key reference to the vector system. For example, in Figure 8.45, Parcel is a table with attributes, linked to the spatial objects stored in a GIS by the Location column. Obviously, several tables may make references to the vector system, but it is not uncommon to have some main table for which the ID is actually also the key.
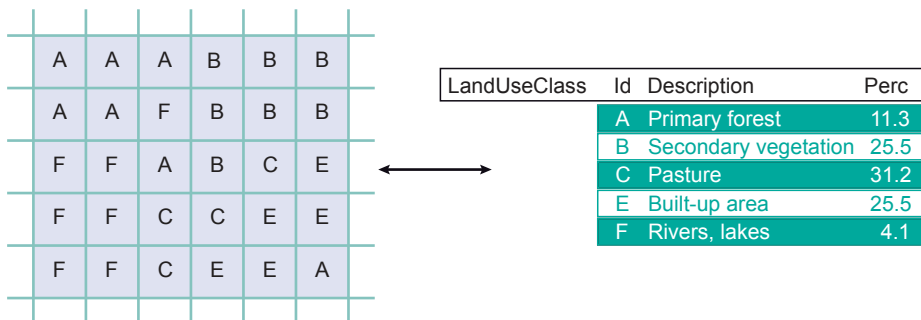
object ID

linking objects and tables

LandUseClass	Id	Description	Perc
	A	Primary forest	11.3
	B	Secondary vegetation	25.5
	C	Pasture	31.2
	E	Built-up area	25.5
	F	Rivers, lakes	4.1

Figure 8.44
A raster representing land use and a related table providing full text descriptions (amongst other things) of each land use class.

Spatial database functionality

DBMS vendors have over the last 20 years recognized the need for storing more complex data, such as spatial data. The main problem was that additional functionality was needed by DBMSs in order to process and manage spatial data. As the capabilities of computer hardware to process information have increased, so too has the desire for better ways of representing and managing spatial data. During the 1990s, object-oriented and object-relational data models were developed for just this purpose. These extend standard relational models by providing support for objects, including "spatial" ones.

Currently, GIS software packages are able to store spatial data using a range of commercial and open-source DBMSs (e.g. Oracle, Informix, IBM DB2, Sybase, and Post-Gres) with the help of spatial extensions. Some GIS software have integrated database "engines" and therefore do not need these extensions. ESRI's ArcGIS, for example, has the main components of the Microsoft Access database software built in. This means that a designer of a GIS application can choose whether to store the application data in the GIS or in the DBMS.

integrated database "engines"

DBMS spatial extension

Spatial databases, also known as geo-databases, are implemented directly on existing

Spatial data

Object ID

Parcel	PId	Location	OwnerID
	3421	2001	435
	8871	1462	550
	2109	2323	1040
	1515	2003	245
	3434	2020	486
	6371	1802	950
	2209	3542	1840
	1505	2609	145
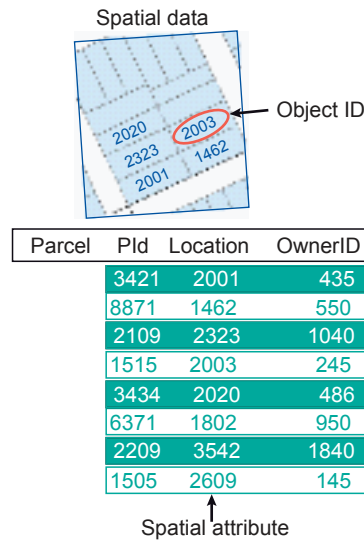
Spatial attribute

Figure 8.45
Storage and linking of vector attribute data between a GIS and a DBMS.

DBMSs using extension software to allow them to handle spatial objects.

There are several advantages in doing this, as we will see below. Put simply, spatial data can be stored in a special database column, referred to as the "geometry" or "feature" or "shape data type", depending on the specific software package. This means GISs can rely fully on a DBMS support for spatial data, making use of a DBMS for data query and storage (and multi-user support), and a GIS for spatial functionality. Small-scale GIS applications may not require a multi-user capability; these applications could be supported by spatial data support from a personal database.

Parcel	PId	Geometry	OwnerID
	3421	"MULTIPOLYGON(((257462.704979333 464780.750851061,257463.89798...)))"	435
	8871	"MULTIPOLYGON(((257409.813950544 464789.91585049,257407.896903...)))"	550
	2109	"MULTIPOLYGON(((257785.714911912 464796.839972167,257782.59794...)))"	1040
	1515	"MULTIPOLYGON(((257790.672100448 464807.13792585,257788.608078...)))"	245
	3434	"MULTIPOLYGON(((257435.527950478 464803.92887633,257428.254887...)))"	486
	6371	"MULTIPOLYGON(((257432.476077854 464813.848852072,257433.147910...)))"	950
	2209	"MULTIPOLYGON(((257444.888027332 464826.555046319,257446.43201...)))"	1840
	1505	"MULTIPOLYGON(((256293.760107491 464935.203846095,256292. 00881...)))"	145

Figure 8.46
Geometry data stored directly in a spatial database table.

A spatial database allows a wide variety of users to access large data sets (both geographic and alphanumeric) and manage their relations, while guaranteeing their integrity. The Open Geospatial Consortium (OGC) has released a series of standards for geographic data formats that (among other things), define:

- which tables must be present in a geo-database (i.e. a geometry columns table and a spatial reference system table);

- the data formats, called "Simple Features" (i.e. point, line, polygon, etc.);

- a set of SQL-like instructions for geographic analysis.

The architecture of a spatial database differs from a standard relational DBMS not only because it can handle geometry data and manage projections, but also because

of the availability a larger set of commands that extend the standard SQL language (distance calculations, buffers, overlay, conversion between coordinate systems, etc.). A geo-database must provide a link between the spatial data represented by rasters or vectors and their non-spatial attribute data.

The capabilities of spatial databases will continue to evolve over time. Currently, ESRI's ArcGIS "Geodatabase" can store topological relationships directly in the database, providing support for different kinds of features (objects) and their behaviour (relations with other objects), as well as ways to validate these relations and behaviours. Effectively, this is the same type of functionality offered by traditional DBMSs, but with geospatial data. Currently, some spatial database packages, such as PostGIS, have full 3D support, as opposed to the 2D support offered by many.

<div style="text-align: right; color: teal;">storing topology</div>

Querying a spatial database   A spatial DBMS provides support for geographic coordinate systems and transformations. It will also provide storage of the relationships between features, including the creation and storage of topological relationships. As a result, one is able to use functions for "spatial query" (exploring spatial relationships). To illustrate, a spatial query using SQL to find all the Thai restaurants within 2 km of a given hotel would look like:

<div style="text-align: right; color: teal;">spatial query</div>

SELECT R.Name
FROM Restaurants AS R,
Hotels as H
WHERE R.Type = Thai AND
H.name = Hilton AND
Intersect(R.Geometry, Buffer(H.Geometry, 2))

The Intersect command creates a spatial join between restaurants and hotels. The Geometry column carries the spatial data. It is likely that in the near future all spatial data will be stored directly in spatial databases.

8.5 GIS Working environment

8.5.1 Spatial Data Infrastructure (SDI)

The way in which spatial data are perceived, expected, and consumed by users in their applications depends, to a large extent, on the current context and shape of technology, projects and markets. Interactions between these three drivers form the basis for the requirements of geoinformation systems at any given time. At present, these interactions translate into systems having to operate in an interconnected environment.

As the systems that rely on spatial data have moved from single, separate working environments towards connected and cooperative environments, different needs, requirements and challenges have emerged. To address these changes, the spatial information community came up with the Spatial Data Infrastructure (SDI)initiative. In [80] an SDI is defined as *the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data*. Several definitions of an SDI exist, however, each adjusted slightly to fit specific needs (see [80]). Regardless of the author or the context, the issue comes down to one objective: interoperability, i.e. the property of diverse systems and organizations that allows them to work together, to inter-operate. The targeted objective of an SDI is, therefore, seamless access to all the constituent elements of a geoinformation system: data, operations, and results. These three elements are collectively called "geo-resources". "Seamless" here means transparently over a network, regardless of computer platform, format or application. Central to this objective are standards.

An SDI is not an entity in itself; it is rather an approach for working efficiently and effectively in a distributed, cooperative environment. There is, therefore, no recipe for the implementation of an SDI. Through the years, experts have come up with different interpretations of the concept and, therefore, different SDI implementations have been created too. The most familiar approach for implementation is based on the notion of a clearinghouse: i.e.ėpository to store descriptions of existing spatial data. These descriptions, known as meta-data, are created and stored in a standardized format. A clearinghouse allows spatial data producers to publish and disseminate meta-data, which in turn can be queried by users to discover spatial data resources. This approach describes the first generation of SDIs, and it was the way to go about implementing them in the early nineties. This could be achieved given the standards available and the maturity of the geoinformation technology of the day. The latest generation of SDI implementations focuses on geo-services. It is based on sounder standards and more robust technology. It uses webservices as a mechanism to provide access to geo-resources. The following subsections describe the developments that are considered to be state of the art in the realm of SDI.

8.5.2 Standards

The underlying working principle of an SDI based on webservices is that it operates on the World Wide Web, also known as the Web. Terms like the Internet and the World Wide Web are often used interchangeably, however they are not one and the same. The Internet is a network, or rather a global system of interconnected *computer networks* that use the standardized Internet Protocol Suite (TCP/IP) to carry a wide range of information resources and services. The most well known application built on top of the Internet is the Web. The Web is a system of interlinked *documents* connected by means of hyperlinks and accessible via the Internet. Other internet-based applications include electronic mail, file transfer, social networking, and multi-player gaming.

In line with this working principle, developers of SDIs have to adhere to two sets of technical standards. The first is the set of technical specifications and guidelines on

which the Web is based. The second is the set of technical specifications that address interoperability issues among geo-resources. The standards for the Web are developed by the World Wide Web Consortium (W3C) [85]. The W3C is an international community, led by Web inventor Tim Berners-Lee, that develops the standards needed to ensure the long-term growth of the Web. Standards for interoperability of geo-resources address a multitude of issues ranging from data capture to presentation and are developed by different organizations, the most of which prominent are the International Organization for Standardization (ISO), through the technical committee ISO/TC 211 [48], and the Open Geospatial Consortium (OGC) [85].
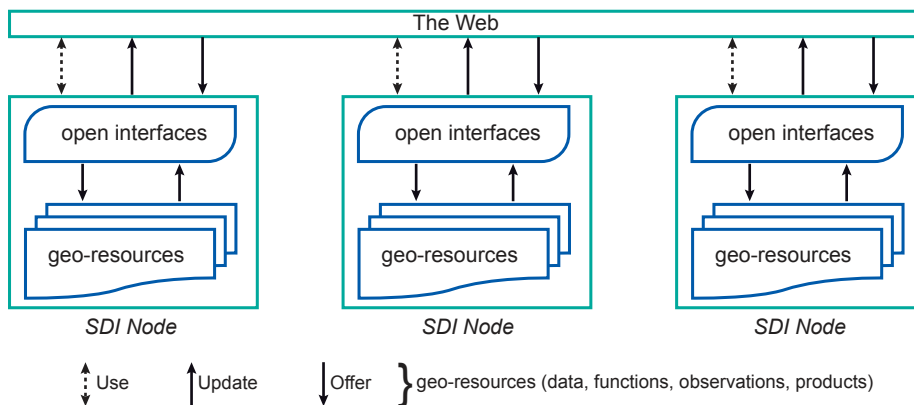
Figure 8.47
Schematic representation of an SDI.

The role of standards in an SDI should only be to support its main objective, i.e. interoperability. In this context, standards are deployed at the interface layer between available geo-resources and their users. Standards that go beyond this purpose, for example data models, should only be used as reference standards in SDI development. SDI participants need to understand the benefits and limitations of this distinction.

Following this paradigm, OGC's standardization efforts provide a comprehensive suite of open interface specifications. An interface is defined as a connection point where two separate system components interact to exchange parameters and instructions. An interface is presented as an ordered set of parameters (with specific names and data types) and instructions (with specific names and functions) for components to interact. An interface that allows any two arbitrary systems (system components) to interact is known as an open interface. Systems that use this type of interface are considered open systems. As a result, an open system is permeable to its environment and can produce a large (potentially infinite) set of services as a result of its interactions. In contrast, a closed system delivers a limited number of services and its sources are only those contained within its boundaries. The relevant OGC standards and how they are used within an SDI are explained in section 8.5.4.

8.5.3 SDI architecture

From an architectural point of view, an SDI is a collaborative network of disparate systems called SDI nodes. An SDI node is a moderately to highly complex information system, usually of long life expectancy, in which geo-resources feature rather prominently. SDI nodes are totally independent systems, which means that they also have to fulfill requirements that are different from those of the SDI itself.

We can argue that we have an SDI in place when we have SDI nodes that host geo-resources that can be found, used and maintained, or even created, by other SDI nodes. Figure 8.48 depicts an schematic view on the components of such an SDI. For the design of an SDI that fulfills the criteria mentioned above, three perspectives need to be

addressed. First, the vertical perspective that focuses on the design of SDI nodes as state-and-function systems that hold data and provide functions in the form of services for operating on that data. Second, the horizontal perspective that focuses on how to specify the communication patterns between SDI nodes, abstractly constructing a larger system, i.e. the SDI. And third, the introspective perspective that focuses on augmenting the specified SDI system with the means to query itself for service possibilities and to allow the dynamic creation of new services.
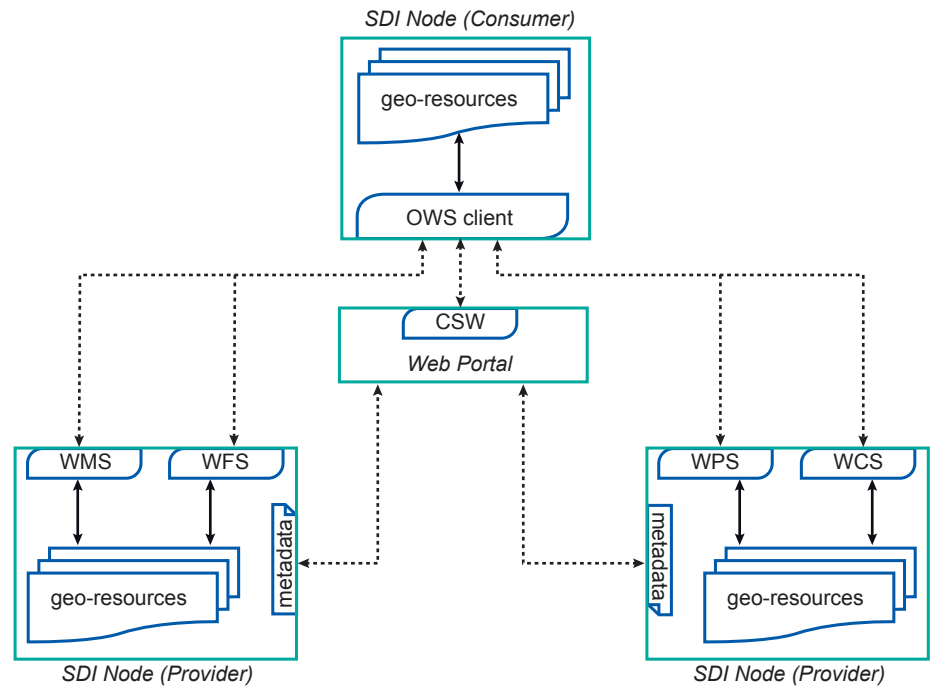
Figure 8.48
Webservices-based SDI architecture.

SDI nodes work based on the well-known client–server architecture, where a partition of responsibilities is enforced. The clients make requests to the server; the server processes the requests and returns the results to the client. In the context of SDI, these roles are interchangeable between SDI nodes. This means an SDI node can play both roles, i.e. that of client or that of server, depending on the circumstances.

8.5.4 Geo-webservices

Modern SDIs are based on webservices, which are defined as mechanisms designed to support interoperable machine-to-machine interaction over the Web. Webservices operate on the basis of standardized technologies and formats/protocols as defined by the W3C (see [126]). Webservices are self-contained, modular applications that can be described, published, located, and invoked over the Web. The working framework on which webservices operate is known as the publish-find-bind paradigm, whereby service providers make services available to service users, who consume resources by locating and binding to services. Interactions between service users and service providers are realized by exchanging messages. These messages are encoded using the eXtensible Markup Language (XML). XML is a general-purpose specification for encoding documents electronically.

Taking advantage of the webservice framework, OGC has developed a set of technical specifications known as OGC Web Services (OWS). OWS specifications are defined using open, non-proprietary Web standards. OWS specifications are platform-neutral

specifications for the implementation of interfaces. Some of the OGC Web Service standards that are relevant for an SDI include (see Figure 8.48):

- The Catalogue Service for the Web (CSW) defines common interfaces to discover, browse, and query meta-data about data, services and other geo-resources. Catalogue services consume and deliver meta-data according to the ISO standards for meta-data. This includes the ISO 19115:2003 *Geographic information–Meta-data*, which defines the schema for the identification, extent, quality, spatial and temporal schema, spatial reference, and distribution of spatial data; and also ISO 19119:2005 *Geographic information–Services*, which defines service meta-data in terms of the architecture patterns for service interfaces used and defines the standard relationship to the Open Systems Environment model. The ISO 19119 standard also presents a taxonomy for services and prescribes how to create a platform-neutral service specification.

- The Web Map Service (WMS) Implementation supports the creation and display of registered and superimposed map-like views of spatial data that come from one source or simultaneously from multiple remote and heterogeneous sources.

- The Web Feature Service (WFS) Implementation Specification allows a client to retrieve and update spatial data encoded in Geography Markup Language (GML) from one or multiple sources. The specification defines interfaces for spatial data access and manipulation operations. Through these interfaces, a Web user or service can combine, use and manage spatial data from different sources. In addition, the transactional version of the WFS specification includes the option to insert, update, or delete features from a vector data source.

- The Web Coverage Service (WCS) Implementation Specification allows clients to access parts of a grid coverage offered by a server. The data served by a WCS are grid data that are usually encoded in a binary image format. The output includes coverage meta-data.

- The Web Processing Service (WPS) Implementation Specification defines an interface that facilitates the publishing of processing functionality and also the discovery of and binding to those processes by clients. A WPS may offer calculations as simple as a buffer, or as complicated as a global climate change model. The data required by a WPS can be delivered across a network using OGC Web Services.

- The Sensor Web Enablement (SWE) set of specifications enable all types of Web and/or Internet-accessible sensors, instruments, and imaging devices to be accessible and, where applicable, controllable via the Web.

Besides the above mentioned interface specifications, an important OGC standard to achieve interoperability is the Geography Markup Language (GML). Its encoding standard is an XML grammar for expressing spatial features. GML serves as a modelling language for geographic systems, as well as an open interchange format for transactions on spatial data over the Web. In keeping with most XML-based grammars, there are two parts to a GML document: the schema that describes the structure of the data contained in the document; and the instance document that contains the actual data.

In addition to the open standards defined by OGC, there are other types of geo-webservices that can be exploited in an SDI environment. Companies like Google and Microsoft have created their own set of services to access geo-resources. These

resources include satellite data, routing operations, and so on. These services are commonly accessed via a web browser. However to properly embed those services within other applications, their developers provide what is known as an application programming interface, or API for short (see Figure 8.49). An API is a set of routines, data structures, object classes and/or protocols that can be used to build applications based on the associated services. An API for the Web is typically defined as a set of request messages along with a definition of the structure of response messages, usually expressed in XML. Most of these commercially-based geo-webservices used Web 2.0 implementation tools (see Subsection 8.5.6).

8.5.5 Meta-data

From a technical point of view, an SDI is a facility that liaises between producers and users of geographic data sets and services. For this arrangement to work, data sets and services have to be made discoverable, analysable and accessible. This is achieved by the creation of descriptions known as meta-data. Meta-data is the mechanism that producers have that enables potential users to find, analyse and evaluate their resources (data sets and services) and determine their fitness for use.

The term meta-data and the meta-data itself have become widely used over the last few years by the geo-community as if it was something new. In reality, however, its underlying concepts have been in use for generations. A map legend, for example, is one embodiment of meta-data, containing details about the publisher of the map, its publication date, the type of map, the spatial reference and the map's scale and accuracy, etc. This connection has become somewhat lost in the transition from analogue to digital data production processes.

Some authors define three categories of meta-data, based on how it is actually used: i.e. discovery, exploration and exploitation meta-data. Discovery meta-data simply enables users to find existing data and services. Discovery meta-data helps answering the question "who has what data/service and from where?", the where being an area of interest defined by means of coordinates, geographical names or administrative areas. Exploration meta-data enable users to determine whether some existing data/service is useful for their application. Exploration meta-data answer questions like "why, when and how was certain data collected". Exploitation meta-data enable users to access, transfer, load, interpret and use data/services in their applications. In addition to access, this type of meta-data also includes details about the price of the data/service and licensing and copyrights.

Meta-data is defined as a formalized and agreed upon set of properties that describe in a significant amount of detail the characteristics of a data set and/or service. ISO has therefore specified in its 19100 suite of standards the set of properties that properly describe a data set and a service.

The ISO 19115 standard defines the meta-data for vector data sets. It is applicable to a whole data set, aggregations of data sets, individual features, and the various classes of objects that compose a feature. This standard defines a large set of meta-data properties (400+), some of which are considered mandatory and some optional. To adhere to the standard, one should implement descriptions that incorporate the mandatory properties and a selection of the optional properties. The result is known as an ISO profile, a subset of the original standard.

As mentioned earlier, meta-data is mainly disseminated via meta-data catalogues that meet the OGC-defined Catalogue Service Web (CSW)-implementation specification. This specification defines the interfaces and binding mechanisms required to publish and access digital catalogues of meta-data for data, services, and related resources. Implementations of the CSW specification are known as Catalogue Services. A repository

of CSW services is known as an OGC Catalogue.

8.5.6 Web portals

Once services have been implemented for different SDI nodes, a common practice is to facilitate their access by building portals. Spatial web portals can be thought of as a gateway that provides access to geo-resources via geo-webservices on the Web. A web portal is simply a website that gives visitors organized access in a unified way, typically through catalog services, to geo-resources on the Web, and preferably also to the people and organizations offering those geo-resources (see Figure 8.49). A portal potentially offers access to many other sites. Consequently, a web portal can also be used to aggregate content.
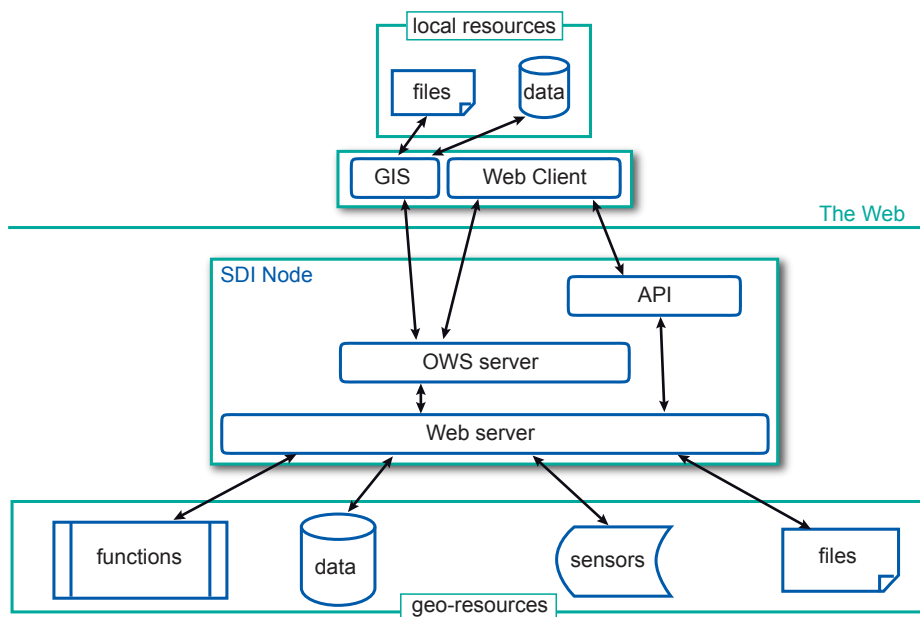
Figure 8.49
SDI node architecture and communication patterns.

8.5.7 Web 2.0

One of the latest developments for the Web, which is also available to SDI developers, is the Web 2.0 concept. Web 2.0 refers to what is perceived as a second generation of Web application development and Web design. Web 2.0 does not refer to any specific change in the technology of the Web, but rather the way in which developers implement websites and thus to the way in which people perceive and use the Web. Web 2.0 applications are characterized by their interactivity and user-centred design. Web 2.0 websites behave similarly to desktop applications that are familiar to computer users. These websites do, therefore, more than just retrieve and display information. Users can exercise control over the activity by using the interactive functions provided by the site. One principle behind Web 2.0 sites is asynchronous communication between the client application and the server. As a consequence, instead of having to reload a webpage whenever there is input from the user, the Web application makes background requests and, based on the response, dynamically updates the sections of the webpage that are affected.

This new way of working on the Web has been adopted by the SDI community and nowadays Web 2.0 tools are available that allow interactive use of spatial data over

the Web. Incidentally most of these tools have come from the open source community. Using these tools, which are available in the form of APIs, SDI developers can build highly-interactive web applications that enable users to perform all sorts of data manipulations over the Web. All the common functions for editing, analysing and processing spatial data that are conventionally only available in desktop GIS packages are now available to users in an SDI environment.

8.6 Data quality

With the advent of satellite remote sensing, GPS and GIS technology, and the increasing availability of digital spatial data, resource managers and others who formerly relied on the surveying and mapping profession to supply high quality map products are now in a position to produce maps themselves. At the same time, GISs are being increasingly used for *decision-support* applications, with increasing reliance on secondary data sourced through data providers or via the internet, from geo-webservices. The consequences of using low-quality data when making important decisions are potentially grave. There is also a danger that uninformed GIS users will introduce errors by incorrectly applying geometric and other transformations to the spatial data held in their database.

application requirements

Below we look at the main issues related to the data quality of spatial data. As outlined in Section 8.1, we will discuss positional, temporal and attribute accuracy, lineage, completeness, and logical consistency. We will begin with a brief discussion of the terms accuracy and precision, as these are often taken to mean the same thing. For a more detailed discussion and advanced topics relating to data quality, the reader is referred to [28].

8.6.1 Accuracy and precision

So far we have used the terms error, accuracy and precision without appropriately defining them. Accuracy should not be confused with *precision*, which is a statement of the smallest unit of measurement to which data can be recorded. In conventional surveying and mapping practice, accuracy and precision are closely related. Instruments with an appropriate precision are employed, and surveying methods chosen, to meet specified tolerances in accuracy. In GISs, however, the numerical precision of computer processing and storage usually exceeds the accuracy of the data. This can give rise to what is known as *spurious accuracy*, for example calculating area sizes to the nearest m^2 from coordinates obtained by digitizing a $1 : 50,000$ map.

accuracy tolerances

The relationship between accuracy and precision can be clarified using graphs that display the probability distribution (see below) of a measurement against the true value T. In Figure 8.50, we depict the cases of good/bad accuracy against good/bad precision.[1] An *accurate* measurement has a mean close to the true value; a *precise* measurement has a sufficiently small variance.

8.6.2 Positional accuracy

The surveying and mapping profession has a long tradition of determining and minimizing errors. This applies particularly to land surveying and photogrammetry, both of which tend to regard positional and height errors as undesirable. Cartographers also strive to reduce geometric and attribute errors in their products, and, in addition, define quality in specifically cartographic terms, for example quality of line work, layout, and clarity of text.

It must be stressed that all measurements made with surveying and photogrammetric instruments are subject to error. These include:

1. Human errors in measurement (e.g. reading errors) generally referred to as gross errors or *blunders*. These are usually large errors resulting from carelessness, which could have been avoided through careful observation, although it is never absolutely certain that all blunders could have been avoided or eliminated.

---

[1]Here we use the terms "good" and "bad" to illustrate the extremes of both accuracy and precision. In real world terms, we refer to whether data are "fit for use" for a given application.
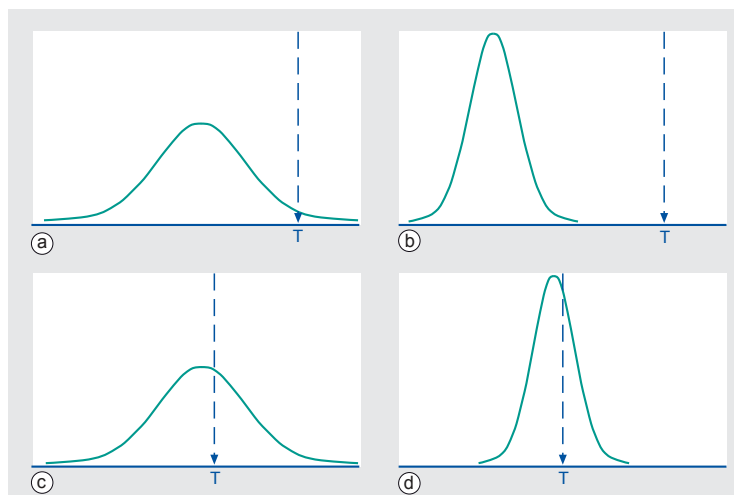
Figure 8.50
A measurement probability function and the underlying true value T: (a) bad accuracy and precision, (b) bad accuracy/good precision, (c) good accuracy/bad precision, and (d) good accuracy and precision.

2. Instrumental or *systematic* errors (e.g. due to maladjustment of instruments). This leads to errors that vary systematically in sign and/or magnitude, but can go undetected by repeating the measurement with the same instrument. Systematic errors are particularly dangerous because they tend to accumulate.

error sources

3. So–called *random* errors caused by natural variations in the quantity being measured. These are effectively the errors that remain after blunders and systematic errors have been removed. They are usually small, and dealt with in least–squares adjustment.

Section 3.2 discussed the errors inherent in various methods of spatial positioning. Below we will at more general ways of quantifying positional accuracy using *root mean square error* (*RMSE*).

Measurement errors are generally described in terms of *accuracy*. In the case of spatial data, accuracy may relate not only to the determination of coordinates (positional error) but also to the measurement of quantitative attribute data. The accuracy of a single measurement can be defined as:

> *"the closeness of observations, computations or estimates to the true values or the values perceived to be true"* [79].

In the case of surveying and mapping, the "truth" is usually taken to be a value obtained from a survey of higher accuracy, for example by comparing photogrammetric measurements with the coordinates and heights of a number of independent check points determined by field survey. Although it is useful for assessing the quality of definite objects, such as cadastral boundaries, this definition clearly has practical difficulties in the case of natural resource mapping where the "truth" itself is uncertain, or boundaries of phenomena become fuzzy. This type of uncertainty in natural resource data is elaborated upon on page 301.

relative and absolute accuracy

Prior to the availability of GPS, resource surveyors working in remote areas sometimes had to be content with ensuring an acceptable degree of *relative accuracy* among the measured positions of points within the surveyed area. If location and elevation are fixed with reference to a network of control points that are assumed to be free of error, then the *absolute accuracy* of the survey can be determined.

Root mean square error

Location accuracy is normally measured as a *root mean square error* (*RMSE*). The RMSE is similar to, but not to be confused with, the standard deviation of a statistical sample. The value of the RMSE is normally calculated from a set of check measurements (co-ordinate values from an independent source of higher accuracy for identical points). The differences at each point can be plotted as error vectors, as is done in Figure 8.51 for a single measurement. The error vector can be seen as having constituents in the x- and y-directions, which can be recombined by vector addition to give the error vector representing the locational error.
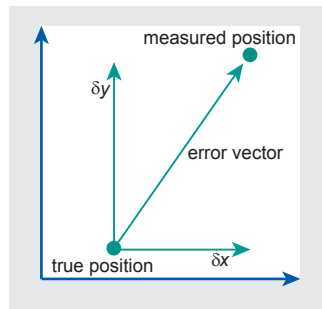
Figure 8.51
The positional error of a measurement can be expressed as a vector, which in turn can be viewed as the vector addition of its constituents in the x- and y-directions, respectively δx and δy.

For each checkpoint, the error vector has components δx and δy. The observed errors should be checked for a *systematic* error component, which may indicate a (possibly repairable) lapse in the measurement method. Systematic error has occurred when $\sum \delta x \neq 0$ or $\sum \delta y \neq 0$.

The systematic error $\delta \overline{x}$ in x is then defined as the average deviation from the true value:

$$\delta \overline{x} = \frac{1}{n} \sum_{i=1}^{n} \delta x_i.$$

Analogously to the calculation of the variance and standard deviation of a statistical sample, the root mean square errors m_x and m_y of a series of coordinate measurements are calculated as the square root of the average squared deviations:

$$m_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \delta x_i^2} \quad \text{and} \quad m_y = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \delta y_i^2},$$

where δx^2 stands for $\delta x \cdot \delta x$. The total RMSE is obtained with the formula

$$m_{\text{total}} = \sqrt{m_x^2 + m_y^2},$$

which, by the Pythagorean rule, is the length of the average (root squared) vector.

Accuracy tolerances

Many kinds of measurement can be naturally represented by a bell-shaped probability density function p, as depicted in Figure 8.52(a). This function is known as the *normal (or Gaussian) distribution* of a continuous, random variable, in the figure indicated as Y. It shape is determined by two parameters: μ, which is the mean expected value for Y, and σ, which is the standard deviation of Y. A small σ leads to a more attenuated bell-shaped function.

distribution of errors

Figure 8.52
(a) Probability density function p of a variable Y, with its mean μ and standard deviation σ. (b) The probability that Y is in the range $[\mu - \sigma, \mu + \sigma]$.
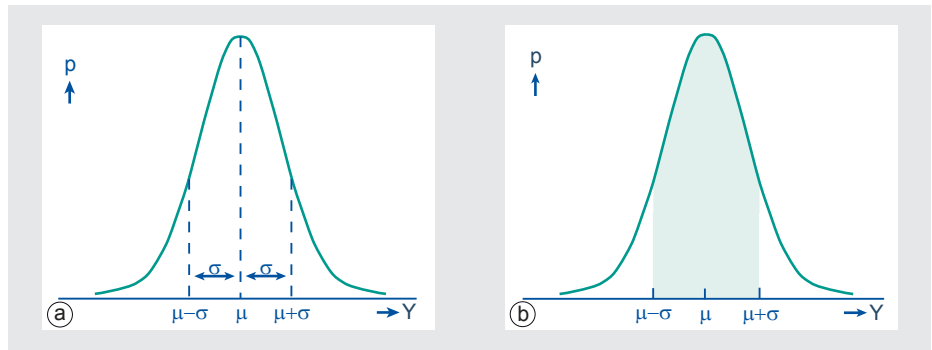
Any probability density function p has the characteristic that the area between its curve and the horizontal axis is equal to 1. Probabilities P can be inferred from p as the area under p's curve. Figure 8.52(b), for instance, depicts $P(\mu - \sigma \leq Y \leq \mu - \sigma)$, i.e. the probability that the value for Y is within distance σ from μ. In a normal distribution this specific probability for Y is always 0.6826.

The *RMSE* can be used to assess the probability that a particular set of measurements does not deviate too much from, i.e. is within a certain range of, the "true" value. In the case of coordinates, the probability density function is often considered to be that of a two-dimensional normally distributed variable (see Figure 8.53). The three standard probability values associated with this distribution are:

- 0.50 for a circle with a radius of $1.1774\ m_x$ around the mean (known as the *circular error probable*, CEP);

- 0.6321 for a circle with a radius of $1.412\ m_x$ around the mean (known as the *root mean square error*, RMSE);

- 0.90 for a circle with a radius of $2.146\ m_x$ around the mean (known as the *circular map accuracy standard*, CMAS).

Figure 8.53
Probability density p of a normally distributed, two-dimensional variable (X, Y) (also known as a normal, bivariate distribution). In the ground plane, starting from the inside out, are the circles associated with CEP, RMSE and CMAS.
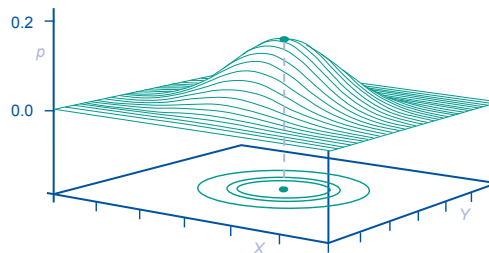
The RMSE provides an estimate of the spread of a series of measurements around their (assumed) "true" values. It is therefore commonly used to assess the quality of transformations such as the absolute orientation of photogrammetric models or the spatial referencing of satellite imagery. The RMSE also forms the basis of various statements for reporting and verifying compliance with defined map accuracy *tolerances*. An example is the American National Map Accuracy Standard, which states that:

> "*No more than 10% of well-defined points on maps of 1:20,000 scale or greater may be in error by more than 1/30 inch.*"

Normally, compliance to this tolerance is based on at least 20 well-defined checkpoints.

The epsilon band

As a line is composed of an infinite number of points, confidence limits can be described by what is known as an epsilon (ε) or Perkal band at a fixed distance on either side of the line (Figure 8.54). The width of the band is based on an estimate of the probable location error of the line, for example to reflect the accuracy of manual digitizing. The epsilon band may be used as a simple means for assessing the likelihood that a point receives the correct attribute value (Figure 8.55).
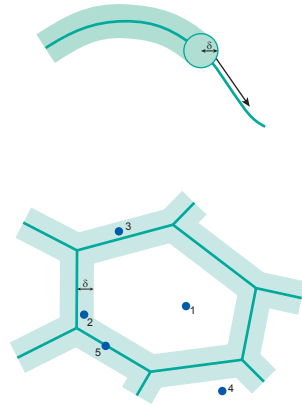
Figure 8.54
The ε or Perkal band is formed by rolling an imaginary circle of a given radius along a line.

Figure 8.55
The ε band may be used to assess the likelihood that a point falls within a particular polygon (source: [86]). Point 3 is less likely part of the middle polygon than point 2.

Describing natural uncertainty in spatial data

There are many situations, particularly in surveys of natural resources, where, according to Burrough, "practical scientists, faced with the problem of dividing up undividable complex continua have often imposed their own crisp structures on the raw data" [13, p. 16]. In practice, the results of classification are normally combined with other categorical layers and continuous field data to identify, for example, areas suitable for a particular land use. In a GIS, this is normally achieved by overlaying the appropriate layers using logical operators.

classification

Particularly in the case of natural resource maps, the boundaries between units may not actually exist as lines but only as transition zones, across which one area continuously merges into another. In these circumstances, rigid measures of positional accuracy, such as RMSE (Figure 8.51), may be virtually insignificant in comparison to the uncertainty inherent in vegetation and soil boundaries, for example.

boundaries

In conventional applications of the error matrix to assess the quality of nominal (categorical) data such as land use, individual samples can be considered in terms of Boolean set theory. The Boolean *membership function* is binary, i.e. an element is either a member of the set (membership is `true`) or it is not a member of the set (membership is `false`). Such a membership notion is well-suited to the description of spatial features such as land parcels for which no ambiguity is involved and an individual ground truth sample can be judged to be either correct or incorrect. As Burrough notes, "increasingly, people are beginning to realize that the fundamental axioms of simple binary logic present limits to the way we think about the world. Not only in everyday situations, but also in formalized thought, it is necessary to be able to deal with concepts that are not necessarily `true` or `false`, but that operate somewhere in between."

membership functions

Since its original development by Zadeh [130], there has been considerable discussion of fuzzy, or continuous, set theory as an approach for handling imprecise spatial data. In GIS, fuzzy set theory appears to have two particular benefits:

fuzzy set theory

1. the ability to handle logical modelling (map overlay) operations on inexact data; and

2. the possibility of using a variety of natural language expressions to qualify uncertainty.

Unlike Boolean sets, fuzzy or continuous sets have a membership function, which can assign to a member any value between 0 and 1 (see Figure 8.56). The membership function of the Boolean set of Figure 8.56a can be defined as MF^B, where:

$$\mathrm{MF}^B(x) = \begin{cases} 1 & \text{if } b_1 \leq x \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$

The crisp and uncertain set membership functions of Figure 8.56 are illustrated for the one-dimensional case. Obviously, in spatial applications of fuzzy set techniques we typically would use two-dimensional sets (and membership functions).

Figure 8.56
(a) Crisp (Boolean) and (b) uncertain (fuzzy) membership functions MF. After Heuvelink [44]
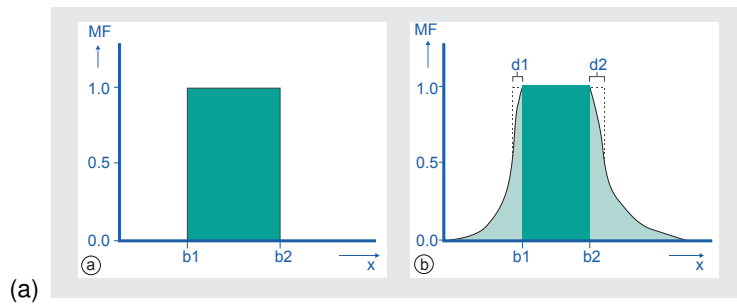
The continuous membership function of Figure 8.56b, in contrast to function MF^B above, can be defined according to Heuvelink [44] as a function MF^C :

$$\mathrm{MF}^C(x) = \begin{cases} \dfrac{1}{1+\left(\frac{x-b_1}{d_1}\right)^2} & \text{if } x < b_1 \\ 1 & \text{if } b_1 \leq x \leq b_2 \\ \dfrac{1}{1+\left(\frac{x-b_2}{d_2}\right)^2} & \text{if } x > b_2 \end{cases}$$

The parameters d_1 and d_2 denote the width of the transition zone around the kernel of the class such that $\mathrm{MF}^C(x) = 0.5$ at the thresholds $b_1 - d_1$ and $b_2 + d_2$, respectively. If d_1 and d_2 are both zero, the function MF^C reduces to MF^B.

An advantage of fuzzy set theory is that it permits the use of natural language to describe uncertainty, for example, "near," "east of" and "about 23 km from". Such natural language expressions can be more faithfully represented by appropriately chosen membership functions.

8.6.3 Attribute accuracy

We can identify two types of attribute accuracies. These relate to the type of data we are dealing with:

- For *nominal or categorical* data, the accuracy of labelling (for example the type of land cover, road surface, etc).

- For *numerical* data, numerical accuracy (such as the concentration of pollutants in a soil, height of trees in forests, etc).

It follows that depending on the data type, assessment of attribute accuracy may range from a simple check on the labelling of features—for example, is a road classified as a metalled road actually surfaced or not?—to complex statistical procedures for assessing the accuracy of numerical data, such as the percentage of pollutants present in a soil.

When spatial data are collected in the field, it is relatively easy to check on the appropriate feature labels. In the case of remotely sensed data, however, considerable effort may be required to assess the accuracy of the classification procedures. This is usually done by means of checks at a number of sample points. The field data are then used to construct an error matrix (also known as a confusion or misclassification matrix) that can be used to evaluate the accuracy of the classification. An example is provided in Table 8.4, where three land use types are identified. For 62 check points that are forest, the classified image identifies them as forest. However, two forest check points are classified in the image as agriculture. *Vice versa*, five agriculture points are classified as forest. Observe that correct classifications are found on the main diagonal of the matrix, which sums up to 92 correctly classified points out of 100 in total.

error matrix

Classified image	Reference data			
	Forest	*Agriculture*	*Urban*	Total
Forest	62	5	0	67
Agriculture	2	18	0	20
Urban	0	1	12	13
Total	64	24	12	100

Table 8.4
Example of a simple error matrix for assessing map attribute accuracy. The overall accuracy is $(62 + 18 + 12)/100 = 92\%$.

8.6.4 Temporal accuracy

As noted, the amount of spatial data sets and archived remotely-sensed data has increased enormously over the last decade. These data can provide useful temporal information, such as changes in land ownership and the monitoring of environmental processes such as deforestation. Analogous to its positional and attribute components, the quality of spatial data may also be assessed in terms of its *temporal accuracy*. For a static feature this refers to the difference in the values of its coordinates at two different times.

Temporal accuracy includes not only the accuracy and precision of time measurements (for example, the date of a survey) but also the temporal consistency of different data sets. Because the positional and attribute components of spatial data may change together or independently, it is also necessary to consider their temporal validity. For example, the boundaries of a land parcel may remain fixed over a period of many years whereas the ownership attribute may change more frequently.

consistency and validity

8.6.5 Lineage

Lineage describes the history of a data set. In the case of published maps, some lineage information may be provided as part of its meta-data, in the form of a note on the data sources and procedures used in the compilation of the data. Examples include the date and scale of aerial photography, and the date of field verification. Especially for digital data sets, however, lineage may be defined more formally as:

> "*that part of the data quality statement that contains information that describes the source of observations or materials, data acquisition and compilation methods, conversions, transformations, analyses and derivations that the data has been*

subjected to, and the assumptions and criteria applied at any stage of its life."
[22]

All of these aspects affect other aspects of quality, for example positional accuracy. Clearly, if no lineage information is available, it is not possible to adequately evaluate the quality of a data set in terms of "fitness for use".

8.6.6 Completeness

Completeness refers to whether there are data lacking in the database compared to what exists in the real world. Essentially, it is important to be able to assess what does and what does not belong to a *complete* data set as intended by its producer. It might be incomplete (i.e. it is "missing" features which exist in the real world), or overcomplete (i.e. it contains "extra" features which do not belong within the scope of the data set as it is defined).

incomplete and overcomplete

Completeness can relate to either spatial, temporal, or thematic aspects of a data set. For example, a data set of property boundaries might be spatially incomplete because it contains only 10 out of 12 suburbs; it might be temporally incomplete because it does not include recently subdivided properties; and it might be thematically overcomplete because it also includes building footprints.

8.6.7 Logical consistency

For any particular application, (predefined) logical rules concern:

- the *compatibility* of data with other data in a data set (e.g. in terms of data format);

- the absence of any *contradictions* within a data set;

- the *topological consistency* of the data set; and

- the allowed attribute *value ranges*, as well as combinations of attributes. For example, attribute values for population, area and population density must agree for all entities in the database.

The absence of any inconsistencies does not necessarily imply that the data are accurate.

8.7 Spatial variation and interpolation

A central activity in studies involving spatial variability is to get from point observations towards area-covering statements. In a GIS one may need a map of a spatial property. Variability (Oxford Dictionary: the quality of being variable in some respect) in space is defined as the phenomenon that a variable changes in space. Earlier, on page 238, we noted that at a certain location a variable may be observed. At a very small distance from this location, the variable may be observed again, and it is likely that it will deviate from the previous observation. The deviations may increase as the distances increase. Description of variability (how large are the deviations if the distance increases) is important for process-based interpretation. An example that will be analysed below concerns the issue of global change, for which weather data is crucial. At a single moment in time, temperature, rainfall and other meteorological data vary in space. Measurements can only be collected reliably at a limited number of locations, making the creation of a map a major undertaking. In another example, to better understand the spatial distribution of wealth and income at a city level, the prices of houses that are for sale can provide important information. Rarely, if ever, will all houses in a city be for sale at a single moment, nevertheless a map of house prices can and should be created. A quantitative approach for making such maps from a collection of point observations within a GIS is called geostatistics. Earth sciences was one of the first disciplines to develop this approach, initially in mining and geology, and later in the agricultural and environmental sciences.

In order to describe spatial variation and interpolation we first single out continuous data. Continuous data represent a continuous phenomenon. They can in principle be collected at any location. Such data are to be distinguished from vector data (roads, buildings, etc.) that one usually considers to be fixed. Notice that house prices are in fact related to a polygon (namely the parcel), but at the city level we can consider them as points. Continuous data are likely to vary throughout a region. Even when all data have been measured in precisely the same manner, i.e. without error and by the same surveyor, variation will still occur. House prices will vary in space, and also the percentage of clay will vary from place to place. Because such variation takes place in space, we speak of spatial variation. Naturally, these variables are allocated to its place in space.

A crucial step in the process is that of obtaining a value at a point where no measurement has been taken. There are several methods for achieving this:

Inverse distance For inverse distance interpolation, weights are assigned to observations. These weights are proportional to the inverse of the distance between a prediction point and an observation point. Distances can be squared although any other power of the distance may also be taken. Inverse distance routines result in a map showing islands, i.e. anomalies in the form of dark and light circles reflecting the values in the observation points.

TIN TIN procedures (discussed on page 246) combine observations by lines. For some data, such as elevation data, TIN procedures are applied successfully. However, TIN procedures fall short if the number of observations is relatively small and if the data (and their locations) are inaccurate.

Trend surfaces Trend surfaces give a global pattern of spatial interpolation. They might give a general picture of the variable in the region.

Geostatistics In essence, in geostatistics the spatial variation of a variable is modelled and then subjected to optimal interpolation. Our focus here is on geostatistics.

For several decades, the field of geostatistics has been exploring approaches for dealing quantitatively with spatial variation in data. Usually, two stages are distinguished. The first stage is an analysis of the (spatial) dependence, i.e. how large is the variation as a function of the distance between observation. The second stage is the joint production of a map of the variable and a map of its precision.

Statistical and geostatistical procedures may be helpful in a number of stages of the interpretation and evaluation of data with a spatial distribution. In many studies they have proven to be indispensable, especially if the amount of available data is great. Some aspects to keep in mind are:

- How can one quantify the *type* and the *amount* of the variation. Some examples: Which properties vary within a region? Does every property varying at the same scale? What is the relation between the spatial variation of a property and aspects of soils, such as sedimentation, and the effect of human activities in the past?

- To *predict* the value of a variable at an unvisited location. Some examples: What is the mean value of nitrate leaching in a parcel? What is the total amount of polluted soil? What is the uncertainty associated with the prediction?

A quantitative approach to spatial variability is of a crucial importance in many spatial studies.

Consider an area which is homogeneous (stationary) with regard to a particular variable being studied. The variogram $\gamma(h)$ is a function of the distance h between locations in the area. The variogram for distance h equals half the expectation of squared differences of variables located at this distance from each other.

8.7.1 The empirical variogram

Observations in space are linked to their coordinates and each observation has its own specific location in space. The value of the coordinate x is essentially linked with the variable Y. Such spatial variables are therefore expressed as $Y(x)$: the place dependence of the variable Y on the location x is given explicitly. They are termed *regionalized variables*. For $Y(x)$ one may read any spatially varying property. The variable $Y(x)$ is put in capitals to indicate that it is a stochastic variable: i.e. a variable that is influenced by unknown and sometimes unmeasurable factors outside our control; it is subject to random influences.

<div style="float:left">regionalized variables</div>

As before, an important characteristic to be dealt with is that the data close to each other are more likely to be similar than data collected at larger distances from each other. This implies that the variables $Y(x_1)$ and $Y(x_2)$ in two locations x_1 and x_2 are probably more alike if the distance between the locations is small, than if the distance is large. The dependence between regionalized variables at different locations is the main, characteristic difference with traditional stochastic variables. The size and the functional form of the differences as a function of the distance will be studied.

An important aspect of regionalized variables is their *expectation* $E[Y(x)] = \mu$ and their *variance* $Var[Y(x)] = \sigma^2$. As the word says: the expectation is the value that would be expected if random influences were absent. For a range of observations, the expectation is estimated by the mean, or by the median (the 50th percentile). The variance is a measure for the noise around the mean: a noisy observation will have a large variance as compared to the expectation, whereas a precise observation would have a low variance. Situations exist, however, where μ does not exist; or σ^2 is not finite. Our examination of geostatistics will focus on the less restrictive requirement,

summarized in what is known as the intrinsic hypothesis [51]. Consider two points along the transect x and $x + h$, the latter point being located at a distance h from the first point x. The intrinsic hypothesis is:

1. $E[Y(x) - Y(x + h)] = 0$

2. $Var[Y(x) - Y(x + h)] < \infty$ and is independent of x.

The first part can be interpreted as follows: the expectation of the difference of a regionalized variable at location x and at a distance h from x equals zero. The second part of the hypothesis requires that the variance of the difference of a regionalized variable measured at location x and at a distance h from x exists, and is *independent of* x. The difference between the two variables $Y(x)$ and $Y(x + h)$ is called a pair difference. The precise form of the dependence of the variance of pair differences on h is often interesting for interpretive purposes, as we will see below.

The spatial dependence function of observations is defined by the second part of the intrinsic hypothesis. It is termed the *variogram* $\gamma(h)$. The variogram is defined as a function of the distance h between locations in the observation space:

variogram

$$\gamma(h) = \frac{1}{2} E\left[Y(x) - Y(x + h)\right]^2 \tag{8.1}$$

Because the expectation of $Y(x) - Y(x + h)$ is equal to zero (intrinsic hypothesis!), the variogram equals half the variance of pair differences at a distance h. Due to the assumption summarized in the intrinsic hypothesis, this variance of pair differences exists and is properly defined. Note that the inclusion of the factor $\frac{1}{2}$ in the expression allows a straightforward comparison with the covariance function: $\gamma(h) = C(0) - C(h)$. The variogram is *independent* of the place where the regionalized variables are located. The squared pair differences have the same expectation, regardless of whether they are measured at one part of the transect or another. Loosely speaking, we expect similar differences between observations independent of the part in the area. In many practical studies one observes an increase of the variogram with increasing distance between the observation locations. This implies that the dependence *decreases* with *increasing* distance h between locations. Loosely speaking: observations close to each other are more likely to be similar than observations at a larger distance from each other.

In order to determine the variogram for data collected in two dimensions, an approach similar to that defined above for the transect data can be applied. The only complication is that for each observation two coordinates are associated, instead of just one coordinate. An estimate, the empirical variogram, $\hat{\gamma}(h)$, may be obtained by taking for a fixed value of h all pairs of points with a separation distance approximately equal to h, squaring the differences between the measurements constituting such a pair, summing these squared differences and dividing the sum by 2 and by the total number of pairs $N(h)$ obtained for this distance. Let the ith pair consist of the two points $y(x_i)$ and $y(x_i + h)$, where we have now used a lower case symbol to distinguish it from the variable $Y(x)$ used earlier. We thus obtain the following equation:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (y(x_i) - y(x_i + h))^2 \tag{8.2}$$

This is repeated for all values of h. A graph may display $\hat{\gamma}(h)$ as a function of h. As the distances between the observation locations are not equal to each other, distance

classes are created: all *pairs* of points of approximately the same sampling distance are grouped into one distance class. This distance defines the class to which the pair belongs.

The user usually has to decide upon a lag length. Sometimes this choice is rather obvious, as in the example of the equidistant observations. A choice for a lag length typically influences the number of distance classes. If the lag length is chosen to be large (larger than the largest occurring distance), only one distance class remains that contains all pair of observations. The estimated variogram value for this distance class equals the estimated variance of the variable: the spatial dependence between the observations is then neglected. At the other extreme, we may choose a very small lag length, resulting in a large number of distance classes, each containing only a few pairs of observations. Although this may be illustrative for some purposes, it is not usually very informative.

For practical purposes there are some general rules that have to be obeyed in order to obtain reliable variogram estimates:

1. The number of pairs of observation points in each class must exceed 30.

2. The maximum distance h between observation points for which the variogram may be determined should not exceed half the length of the area.

In some programs a lag tolerance also has to be specified: if the lag tolerance is less than half the lag length, pairs of observations may be excluded from the analysis; whereas if it exceeds half the lag length, pairs of observations may be allocated to different distance classes.

Figure 8.57
Four common variogram models, each with a nugget equal to 1, a sill equal to 6, and a range equal to 4. The sill is only reached by the spherical model, the other models are within 95% of the sill when the distance is equal to the range.

It is often necessary to fit a specific function through the variogram estimates. A practical way to do this is to estimate the parameters of such a function by a non-linear regression procedure. A distinction can be made between transitive variograms (which apply to fields that have a finite variance) and infinite variograms. Commonly

used transitive variograms are the spherical variogram, the exponential variogram, the Gaussian variogram and the hole effect (or wave) variogram, see Figure 8.57. These variograms all depend upon the distance h. They are characterized by two parameters, a range of influence a and the sill variance b. The *range a* is a measure for the distance up to which the spatial dependence extends. Between locations separated by a distance exceeding a, the regionalized variables are uncorrelated; between locations separated by a distance smaller than a the regionalized variables are dependent. The *sill value* (or the variance) is the value b that the variogram reaches if h tends to infinity, i.e. if the observations are growing to be uncorrelated. A special variogram is the Nugget variogram, which takes a constant value (C_0) for all distances h. The *nugget effect*, a term borrowed from gold mining, denotes the non-spatial variability, the variability at very small distances or the operator bias. If a sill value and a range are not observed then an infinite variogram is the most appropriate. The most common infinite variogram is the power variogram that is characterized by two parameters: the power m and a multiplication parameter k.

range

sill

nugget

Any sum of variograms can be made, in particular the nugget effect is often added to other variograms, resulting in a variogram with a discontinuity at the origin. By definition, $\gamma(0) = 0$. In all the equations, C_0, A and b, or k and m, are positive parameters that are to be determined from the original data. The exponential variogram never reaches the sill value (nor does the Gaussian variogram or the Hole effect variogram). The parameter a is therefore associated with the range but is not similar to the range. We define the effective range to be equal to $3 \cdot a$, being the distance where the exponential model reaches 95% of the sill value. Similar values apply for the other variograms. The Gaussian variogram is characterized by its horizontal behaviour at the origin. This variogram is encountered, for example, when there is uncertainty with respect to the precise location of the observations. The hole effect (wave) variogram is regularly encountered in practice. It points to periodicities of the variable caused by human influences, sedimentation processes, etc. Interpretation of such periodicities is often important.

Whenever a sill value is reached it can be interesting to study the sill/nugget ratio, which gives an indication of the part of the variability to be assigned to spatial variability and of the part to be assigned to non-spatial variability. If the ratio is close to 1, the non-spatial variability is dominant, otherwise the spatial variability is.

8.7.2   Interpolation

Interpolation is used to create a GIS layer out of point observations on a continuous variable. The reason for doing this could be manifold: for visualization purposes, for making a proper reference with other data, or for making a combination of different layers. Consider the problem of obtaining $Y(x_0)$, i.e. the value of a variable $Y(x)$ at an unvisited location x_0. A basic fact is Tobler's law (see Subsection 8.1.2). This already implies intuitively that it is highly unlikely that all the predictions are equal to the mean value: deviations from the mean are likely to occur, especially in the neighbourhood of largely deviating observations. If an observation is above the regression line, then it is highly likely that observations in its neighbourhood will be above this line as well.

We will focus attention on linear combinations of the observations that we will call a predictor for $Y(x_0)$. Hence, each observation is assigned a *weight* such that the predictor is without bias (i.e. the predictor may yield too high a value or too low a value, but on the average it is just right). Carrying out predictions is never precise, and the difference between the value of the variable had we observed it and its interpolated value is called the prediction error. We are interested in predictions that have the lowest

variance of the prediction error.

Let the optimal predictor be denoted with $\hat{Y}(x_0)$. It is linear in the observations, therefore

$$\hat{Y}(x_0) = \sum_{i=1}^{n} \lambda_i Y(x_i) \tag{8.3}$$

with as yet unknown weight λ_i. For the prediction error $e = \hat{Y}(x_0) - Y(x_0)$ it is assumed that its expectation is zero ($E[e] = 0$) and that its variance is minimal among all linear unbiased predictors: $Var[\hat{Y}(x_0) - Y(x_0)]$ is minimal. The variance of the prediction error is of importance and is to be calculated below; see equation 8.5.

Predicting requires knowledge of the variogram. Suppose, therefore, that there are n observations, that a variogram has been determined, that a model has been fitted, and that its parameters have been estimated. The variogram values can be determined for all distances between all pairs of points consisting of observation points (which are $\frac{1}{2}n(n-1)$ in number). They are contained in the, symmetric, $n \times n$ matrix G. The elements of G, g_{ij}, are then filled with the values obtained from $g_{ij} = \gamma(|x_i - x_j|)$; g_{ii} contains the variogram for the distance between x_i and x_i, a pair of points with distance equal to zero, and hence $g_{ii} = 0$; g_{ij} for $i \neq j$ contains the variogram value for the distance between x_i and x_j and is equal to g_{ij}. The variogram can also be determined for all pairs of points consisting of an observation point and the prediction location (which are n in number). These are contained in the vector g_0. The ith element of g_0 contains the variogram value for the distance between x_i and the prediction location x_0. We first estimate the spatial mean $\hat{\mu}$ by means of the generalized least squares estimator

$$\hat{\mu} = \left(1_n^T G^{-1} 1_n\right)^{-1} 1_n^T G^{-1} y \tag{8.4}$$

where y is the vector of observations and 1_n is the vector of n elements, all equal to 1, and T denotes the transpose of a vector (or matrix). Notice that it is different from the average value, as it essentially includes the spatial dependence.

Equation 8.3 can be further modified as

$$\hat{Y}(x_0) = \sum_{i=1}^{n} \lambda_i Y(x_i) = \hat{\mu} + g_0^T G^{-1} (y - \hat{\mu} \cdot 1_n) \tag{8.5}$$

The equation consists of two terms: the spatial mean $\hat{\mu}$ and the term $g_0^T G^{-1}(y - \hat{\mu} \cdot 1_n)$. This second term expresses the influence on the predictor of the residuals $(y - \hat{\mu} 1_n)$ of the observations y with respect to the mean value $\hat{\mu} 1_n$. The residuals are transformed with $g_0^T G^{-1}$. Such a transformation is clearly based upon the variogram. Variogram values among the observation points are included in the matrix G, variogram values between the observation points and the prediction location in the vector g_0. Predicting an observation in the presence of spatially dependent observations is termed Kriging, named after the first practitioner of these procedures, the South African mining engineer Daan Krige, who did much of his early empirical work in the Witwatersrand gold mines.

Kriging

Every prediction is associated with a prediction error. The prediction error itself cannot be determined, but an equation for the variance of the prediction error is given by

prediction error variance

$$Var(Y(x_0) - \hat{Y}(x_0)) = -g_0^T G^{-1} g_0 + \frac{x_a^2}{V} \qquad (8.6)$$

where x_a equals $1 - g_0^T G^{-1} 1_n$ and $V = 1_n' G^{-1} 1_n$. All matrices and vectors can be filled on the basis of the data, the observation locations and the estimated variogram. We remark that the variance of the prediction error depends only *indirectly* upon the observations: the vector y is not included in the equation. However, the variogram is estimated on the basis of the observations, and hence the observations appear in an indirect way in the equation. The *configuration* of the n observation points and the one prediction location influences the prediction error variance as well.

If a map has to be constructed of a spatial property for which the observations are collected in a 2-dimensional space the following procedure may be used:

1. Determine the empirical variogram;

2. Fit a variogram to the empirical variogram;

3. Predict values at the nodes of a fine-meshed grid;

4. Present the results in a two- or a three-dimensional perspective by linking individual predictions with line elements.

In addition to the map itself, it may be desirable (and sometimes even necessary) to display the prediction error variance (or its square root), which is obtained at the same nodes of the fine-meshed grid as the predictions themselves. This map displays the spatial uncertainty of the map.

In the manner described above, it is possible to jointly make two layers in a GIS by spatial interpolation of point observations. Such interpolation has the property that it is driven by the spatial variability of the continuous variable and is, hence, specific for each variable.
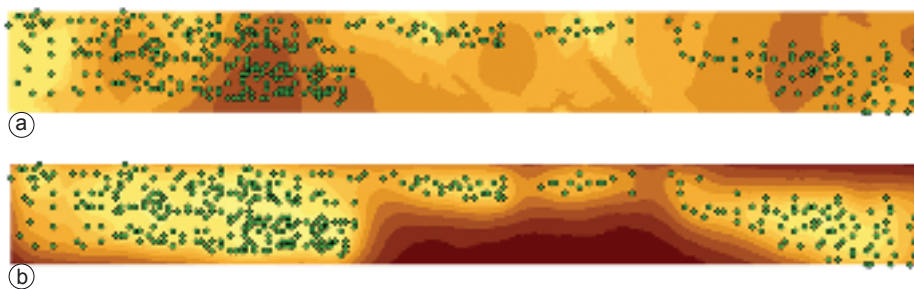
Figure 8.58
Ordinary kriging of lead (Pb) concentrations in soil within an area in The Hague. x-coordinate ranges over 2 km, y-coordinate ranges over 200 m. (a) predictions of the concentrations (values from 6 to 2900 ppm), (b) kriging standard deviations (values from 316 to 359).

To illustrate that we consider an example. In the city of The Hague a soil inventory was carried out within an area of 2 km by 200 m (Figure 8.58), bordering a railway. This inventory followed the closing of a cable factory in the area, and the decision making on the intended future use of the area. In total some 500 observations were provided on several chemical constituents of the possibly contaminated soil. One of the critical constituents is the amount of lead (Pb) that was sampled from the top 50 cm of the soil. In order to get a full overview of the spatial variation of this variable, a variogram was constructed and two maps were produced: the kriged map of the Pb contamination in the soil (Figure 8.58a), and a map of the standard deviation of the prediction error (Figure 8.58b). Maps were produced with ArcGIS. The top map showed a large spatial

variation, from peak values up to 2900 ppm (parts per million, equivalent to 2.9 g Pb per kg soil), and critical environmental thresholds were exceeded. Such heavily polluted soils have to be cleaned, in particular if the future use of the area would be residential. The kriging standard deviation showed less variation in absolute values, but it shows relatively low values close to observation locations, and increasing values at locations that are farther away from the sampling locations.

Chapter 9

Analysis and Process modelling

Rolf de By
Otto Huisman
Menno-Jan Kraak

Introduction

We know from preceding chapters that the analytical capabilities of a GIS make use of spatial and non-spatial (attribute) data to answer questions and solve problems that are of spatial relevance. We now make a distinction between analysis (or analytical operations) and analytical models (often referred to as "modelling"). And by analysis we actually mean only a subset of what is usually implied by the term: we do not specifically deal with advanced statistical analysis (such as cluster detection or geostatistics), which is beyond the scope of this textbook.

All knowledge of the world is based on models of some kind—whether idiosyncratic interpretations, culturally-based stereotypes or complex equations that describe a physical phenomenon. We have already seen in Chapter 7 that there are different types of models and that the word itself has different meanings in different contexts. Section 8.1 notes that even spatial data are themselves a kind of "model" of some part of the real world. Here, in this chapter, we focus on analytical functions that form the building blocks for application models. Section 9.1 presents a framework that defines four main classes of these functions. Each class is then discussed, with the help of examples, in detail in subsequent sections. The aim is to make clear to the reader that these operations can be combined in various ways to perform increasingly complex analyses.

9.1 Classification of analytical GIS capabilities

There are many ways to classify the analytical functions of a GIS. The classification presented here is essentially the one put forward by Aronoff [4], which makes the following distinctions:

- Classification functions allow the assignment of features to a class on the basis of attribute values or attribute ranges (definition of data patterns). On the basis of reflectance characteristics found in a raster, pixels may be classified as representing different crops, e.g. potato or maize.

- Retrieval functions allow selective searching of data. We might, for example, retrieve all agricultural fields on which potato is grown.

- Generalization functions allow different classes of objects with common characteristics to be joined to form a higher-level (generalized) class. For example, we might generalize fields where potato or maize, and possibly other crops, are grown as "food-produce fields".

- Measurement functions allow the calculation of distances, lengths or areas. All functions in this category are performed on a single (vector or raster) data layer, often using the associated attribute data.

More detail can be found in Section 9.2. The following three function types belong to the above classification functions.

Overlay functions

Overlay functions is one of the most frequently used functions in a GIS application. They combine two (or more) spatial data layers, comparing them position by position and treating areas of overlap—and of non-overlap—in distinct ways. Many GISs support overlays through an algebraic language, expressing an overlay function as a formula in which the data layers are the arguments. In this way, we can find:

- those potato fields on clay soils (select the "potato" cover in the crop-data layer and the "clay" cover in the soil-data layer and perform an intersection of the two areas found);

- those fields in which potato or maize is the crop (select both areas of "potato" and "maize" cover in the crop-data layer and determine their union);

- those potato fields not on clay soils (use a difference operator of areas with "potato" cover with the areas having clay soil);

- those fields that do not have potato as a crop (determine the complement of the potato areas).

Examples are provided in Section 9.3.

Neighbourhood functions

Neighbourhood functions evaluate the characteristics of an area surrounding a feature's location. A neighbourhood function "scans" the neighbourhood of the given feature(s), and performs a computation on it(them)

- Search functions allow the retrieval of features that fall within a given search window. This window may be a rectangle, circle or polygon.

- Buffer zone generation (or buffering) is one of the best-known neighbourhood functions. It determines a spatial envelope (buffer) around a given feature or features. The buffer created may have a fixed width or a variable width that depends on characteristics of the area.

- Interpolation functions predict unknown values using the known values at nearby locations. This typically occurs for continuous fields, e.g. elevation, when the data actually stored does not provide a direct answer for the location/locations of interest.

- Topographic functions determine characteristics of an area by also looking at the immediate neighbourhood. Typical examples are slope computations on digital terrain models (i.e. continuous spatial fields). The slope at a location is defined as the plane tangent to the topography at that location. Various computations can be performed, such as:

  - determination of slope angle;
  - determination of slope aspect;
  - determination of slope length;
  - determination of contour lines.

These functions are discussed more fully in Section 9.4.

Connectivity functions

Connectivity functions work on the basis of networks, including road networks, water courses in coastal zones, and communication lines in mobile telephony. These networks represent spatial linkages between features. Main functions of this type include:

- Contiguity functions for evaluating a characteristic of a set of connected spatial units. One can think, here, of the search for a contiguous area of forest of a certain size and shape in a satellite image.

- Network analytic functions for computing related to connected line features that make up a network. The network may consist of roads, public transport routes, high-voltage power lines, or other forms of transportation infrastructure. Analysis of such networks may entail shortest path computations (in terms of distance or travel time) between two points in a network for routing purposes. Other forms are to find all points reachable within a given distance or duration from a start point for allocation purposes, or determination of the capacity of the network for transportation between an indicated source location and sink location.

- Visibility functions also fit in this list because they are used to compute the points visible from a given location (viewshed modelling or viewshed mapping) using a digital terrain model.

For more details see Section 9.5.

9.2 Measurement, retrieval and classification

9.2.1 Measurement

Geometric measurement on spatial features includes counting, distance and area size computations. This subsection discusses such measurements in a planar spatial reference system. We limit ourselves to geometric measurements and do not include attribute data measurement, which is typically performed in a database query language. Measurements on vector data are more advanced (and thus also more complex) than those on raster data.

Measurements on vector data

The primitives of vector data sets are the point, (poly)line and polygon. Related geometric measurements are location, length, distance and area size. Some of these are geometric properties of a feature in isolation (location, length, area size); others (distance) require two features to be identified.

Figure 9.1
The centroid: (a) the centroid of an object can be visualized as the point on which the object would balance when placed on the tip of a pencil, (b) the centroid of a polygon, (c) the centroid of the ITC building is actually outside the building (in the courtyard).
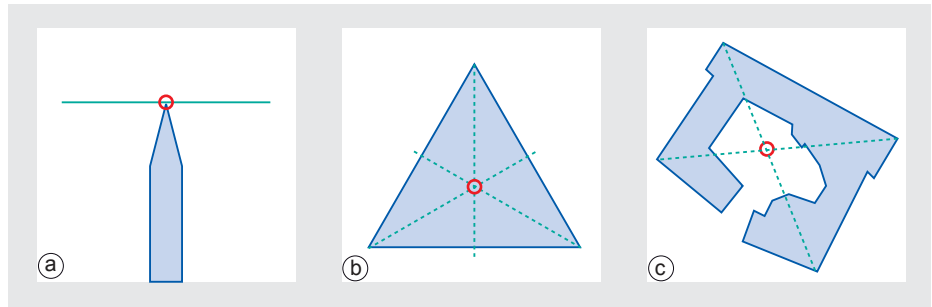
centroid

The location property of a vector feature is always stored by the GIS: a single coordinate pair for a point, or a list of pairs for a polyline or polygon boundary. Occasionally, there is a need to obtain the location of the centroid of a polygon (see some examples in Figure 9.1); some GISs store these also, while others compute them on the fly.

Length is a geometric property associated with polylines, by themselves or in their function as polygon boundaries. It can obviously be computed by the GIS—as the sum of lengths of the constituent line segments—but quite often it is also stored with the polyline.

Area size is associated with polygon features. Again, it can be computed, but it is usually stored with the polygon as an extra attribute value. This speeds up the computation of other functions that require area size values. The attentive reader will have noted that all of the above "measurements" do not actually require computation but only retrieval of stored data.

Measurement of distance between two features is another important function. If both features are points, say p and q, the computation in a Cartesian spatial reference system is given by the well-known Pythagorean distance function:

$$dist(p,\, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

If one of the features is not a point, or both are not, we must be precise in defining what we mean by their distance. All these cases can be summarized as computation of the minimal distance between a location occupied by the first feature and a location occupied by the second feature. This means that features that intersect or meet, or when one contains the other, have a distance of 0. We leave a further case analysis, including polylines and polygons, to the reader as an exercise. It is not possible to store all distance values for all possible combinations of two features in any reasonably-sized spatial database. As a result, the system must compute on-the-fly whenever a distance computation request is made.

minimal bounding box

Another geometric measurement used by GISs is the minimal bounding box computation. It applies to polylines and polygons and determines the minimal rectangle—with sides parallel to the axes of the spatial reference system—that covers the feature. This is illustrated in Figure 9.2. Bounding box computation is an important support function of GISs: for instance, if the bounding boxes of two polygons do not overlap, we know the polygons cannot possibly intersect each other. Since polygon intersection is

a complicated function but bounding box computation is not, a GIS will always first apply the latter as a test to see whether it must do the first.

For practical purposes, it is important to be aware of the unit of measurement that applies to the spatial data layer that one is working on. This is determined by the spatial reference system that has been defined for it during data preparation.
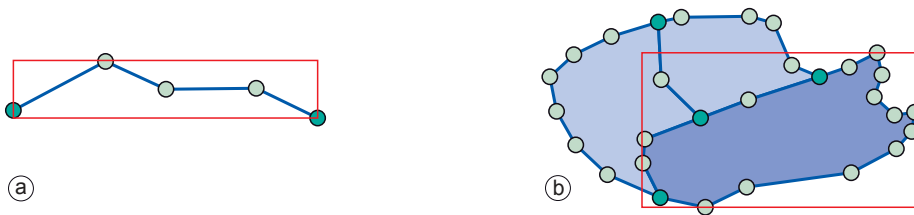
A common use of area size measurements is when one wants to sum up the area sizes of all polygons belonging to some class. This class could be crop type: what is the size of the area covered by potatoes? If our crop classification is in a stored data layer, the computation would include (a) selecting the potato areas, and (b) summing up their (stored) area sizes. Clearly, little geometric computation is required in the case of stored features. This is not the case when we are interactively defining our vector features for GIS use and we want measurements to be performed on these interactively defined features. Then the GIS will have to perform complicated geometric computations.

Measurements on raster data

Measurements on raster data layers are simpler because of the regularity of the cells. The area size of a cell is constant and is determined by the cell resolution. Horizontal and vertical resolution may differ, but typically they do not. Together with the location of what is called an anchor point, this is the only geometric information stored with the raster data, so all other measurements by the GIS are computed. The anchor point is fixed by convention to be the lower-left (or sometimes upper-left) location of the raster.

raster's anchor point

Location of an individual cell derives from the raster's anchor point, the cell resolution, and the position of the cell in the raster. Again, there are two conventions: the cell's location can be its lower-left corner, or the cell's midpoint. These conventions are set by the software in use, and in cases of data of low resolution it becomes more important to be aware of them. The area size of a selected part of the raster (a group of cells) is calculated as the number of cells multiplied by the cell-area size. The distance between two raster cells is the standard distance function applied to the locations of their respective midpoints; obviously the cell resolution has to be taken into account. Where a raster is used to represent line features as strings of cells through the raster, the length of a line feature is computed as the the sum of distances between consecutive cells.

9.2.2 Spatial selection queries

When exploring a spatial data set, the first thing one usually wants to do is select certain features, to (temporarily) restrict the exploration. Such selections can be made on geometric/spatial grounds or on the basis of attribute data associated with the spatial features. We discuss both techniques in the following two subsections.

Interactive spatial selection

interactive selection

In interactive spatial selection, one defines the selection condition by pointing at or drawing spatial objects on the screen display, after having indicated the spatial data layer(s) from which to select features. The interactively defined objects are called the selection objects; they can be points, lines, or polygons. The GIS then selects the features in the indicated data layer(s) that overlap (i.e. intersect, meet, contain, or are contained in; see Figure 8.13) with the selection objects. These become the selected objects.

As we have seen in Section 8.1, spatial data are usually associated with their attribute data (stored in tables) through a key/foreign key link. Selections of features lead, via these links, to selections on the records. Conversely, selection of records may lead to selection of features.

Figure 9.3
All city wards that overlap with the selection object—here a circle—are selected (left), and their corresponding attribute records are highlighted (right; only part of the table is shown). Data from an urban application in Dar es Salaam, Tanzania.
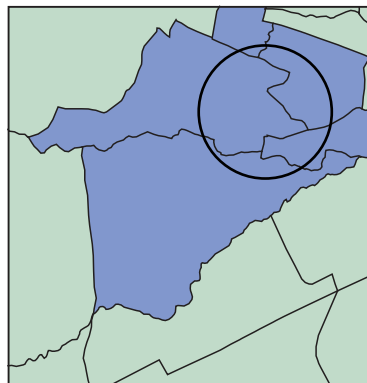
Area	Perimeter	Ward_id	Ward_name	District	Pop88	Pop92
65420380.0000	41654.940000	1	KUNDUCHI	Kinondoni	22106	27212.00
24813620.0000	30755.620000	2	KAWE	Kinondoni	32854	40443.00
18698500.0000	26403.580000	3	MSASANI	Kinondoni	51225	63058.00
81845610.0000	49645.160000	4	UBUNGO	Kinondoni	47281	58203.00
4468546.00000	13480.130000	5	MANZESE	Kinondoni	59467	73204.00
4999599.00000	10356.850000	6	TANDALE	Kinondoni	58357	71837.00
4102218.00000	8951.096000	7	MWANANYAMALA	Kinondoni	72956	89809.00
3749840.00000	9447.420000	8	KINONDONI	Kinondoni	42301	52073.00
2087509.00000	7502.250000	9	UPANGA WEST	Ilala	9852	11428.00
2268513.00000	9028.788000	10	KIVUKONI	Ilala	5391	6254.00
1400024.00000	6883.288000	11	NDUGUMBI	Kinondoni	32548	40067.00
888966.900000	4589.110000	12	MAGOMENI	Kinondoni	16938	20851.00
1448370.00000	5651.958000	13	UPANGA EAST	Ilala	11019	12782.00
6214378.00000	14552.080000	14	MABIBO	Kinondoni	43381	53402.00
2496622.00000	7121.255000	15	MAKURUMILA	Kinondoni	54141	66648.00
1262028.00000	4885.793000	16	MZIMUNI	Kinondoni	23989	29530.00
35362240.0000	28976.090000	17	KINYEREZI	Ilala	3044	3531.00
1010613.00000	5393.771000	18	JANGIWANI	Ilala	15297	17745.00
475745.500000	3043.068000	19	KISUTU	Ilala	8399	9743.00
1754043.00000	7743.187000	20	KIGOGO	Kinondoni	21267	26180.00
29964950.0000	36964.000000	21	KIGAMBONI	Temeke	23203	27658.00
1291479.00000	5187.690000	22	MICHIKICHINI	Ilala	14852	17228.00
720322.100000	4342.732000	23	MCHAFUKOGE	Ilala	8439	9789.00
9296131.00000	16321.530000	24	TABATA	Ilala	18454	21407.00
483620.700000	3304.072000	25	KARIAKOO	Ilala	12506	14507.00
3564653.00000	9586.751000	26	BUGURUNI	Ilala	48286	56012.00
2639575.00000	6970.186000	27	ILALA	Ilala	35372	41032.00
912452.800000	4021.937000	28	GEREZANI	Ilala	7490	8688.00
6735135.00000	13579.590000	29	KURASINI	Temeke	26737	31871.00

Interactive spatial selection answers questions like "What is at ...?" In Figure 9.3, the selection object is a circle and the selected objects are the blue polygons; they overlap with the selection object.

Spatial selection by attribute conditions

attribute selection

It is also possible to select features by using selection conditions on feature attributes. These conditions are formulated in SQL (if the attribute data reside in a relational database) or in a software-specific language (if the data reside in the GIS itself). This type of selection answers questions like "Where are the features with ...?"

Figure 9.4 shows an example of selection by attribute condition. The query expression is Area < 400,000, which can be interpreted as "Select all areas of land use of which the size is less than 400,000." The polygons in red are the selected areas; their associated records are also highlighted in blue.

We can use a set of features already selected as the basis for further selection. For instance, if we are interested in land use areas of size less than 400,000 that are of land use type 80, the selected features of Figure 9.4 are subjected to a further condition, Land Use = 80. The result is illustrated in Figure 9.5.
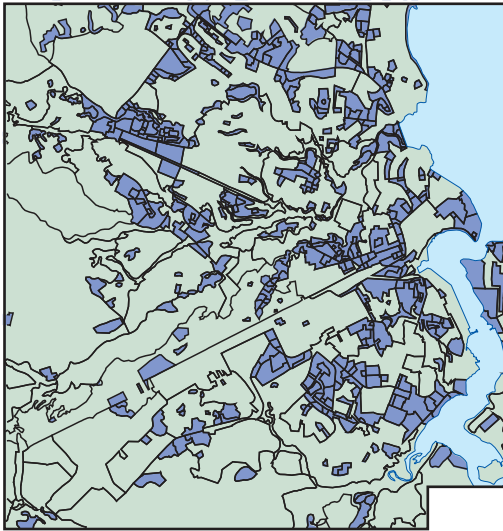
Area	IDs	LandUse
174308.70	2	30
2066475.00	3	70
214582.50	4	80
29313.86	5	80
73328.08	6	80
53303.30	7	80
614530.10	8	20
1637161.00	9	80
156357.40	10	70
59202.20	11	20
83289.59	12	80
225642.20	13	20
28377.33	14	40
228930.30	15	30
986242.30	16	70

Figure 9.4
Spatial selection using the attribute condition Area < 400,000 on land use areas in Dar es Salaam. Spatial features on the left, associated attribute data (in part) on the right.
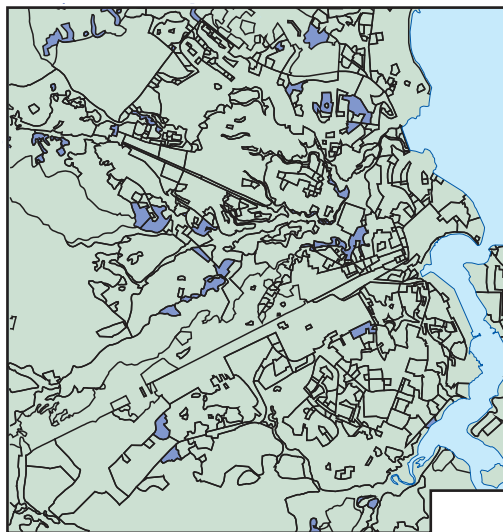
Area	IDs	LandUse
174308.70	2	30
2066475.00	3	70
214582.50	4	80
29313.86	5	80
73328.08	6	80
53303.30	7	80
614530.10	8	20
1637161.00	9	80
156357.40	10	70
59202.20	11	20
83289.59	12	80
225642.20	13	20
28377.33	14	40
228930.30	15	30
986242.30	16	70

Figure 9.5
Further spatial selection from the features already selected in Figure 9.4 using the additional condition Land Use = 80 on land use areas. Note that fewer features are now selected.

Combining attribute conditions

The combination of conditions just dealt with in the previous subsection is fairly common in practice.

When multiple criteria have to be used for selection, we need to carefully express all of these in a single composite condition. The tools for this come from a field of mathematical logic known as propositional calculus.

The example of the previous subsection made use of simple atomic conditions such as Area < 400,000 and Land Use = 80. Atomic conditions use a predicate symbol, such as < (less than) or = (equals). Other possibilities are <= (less than or equal), > (greater than), >= (greater than or equal) and <> (does not equal). Any of these symbols is combined with an expression on the left and one on the right. For instance, Land Use <> 80 can be used to select all areas with a land use class different from 80. Expressions are either constants like 400,000 and 80, attribute names like Area and

Land Use, or possibly composite arithmetic expressions like 0.15×Area, which would compute 15% of the area size.

atomic condition

Atomic conditions can be combined into composite conditions using logical connectives. The most important ones are AND, OR, NOT and the bracket pair (. . .). If we write a composite condition such as Area < 400,000 AND Land Use = 80, we can use it to select areas for which both atomic conditions hold true. This is the meaning of the AND connective. If we had written Area < 400,000 OR Land Use = 80 instead, the condition would have selected areas for which either condition holds, so effectively those with an area size less than 400,000, but also those with land use class 80. (Included, of course, will be areas for which both conditions hold.)

The NOT connective can be used to negate a condition. For instance, the condition NOT (Land Use = 80) would select all areas with a different land use class than 80. (Clearly, the same selection can be obtained by writing Land Use <> 80 but this is not the point.) Finally, brackets can be applied to force grouping amongst atomic parts of a composite condition. For instance, the condition (Area < 30,000 AND Land Use = 70) OR (Area < 400,000 AND Land Use = 80) will select areas of class 70 less than 30,000 in size, as well as class 80 areas less than 400,000 in size.

composite condition

Spatial selection using topological relationships

spatial selection

Various forms of topological relationship between spatial objects were discussed in the subsection in Chapter 8 titled Topology and spatial relationships (page 249). These relationships can be useful to select features as well. The steps to be carried out are:

1. select one or more features as the selection objects; and

2. apply a chosen spatial relationship function to determine the selected features that have that relationship with the selection objects.

Selecting features that are inside selection objects

This type of query uses the containment relationship between spatial objects. Obviously, polygons can contain polygons, lines or points, and lines can contain lines or points, but no other containment relationships are possible.

Figure 9.6a illustrates a containment query. Here, we are interested in finding the location of medical clinics in Ilala District. We first selected all areas of Ilala District using the technique of selection by attribute condition District = "Ilala". Then, these selected areas were used as selection objects to determine which medical clinics (as point objects) were within them.

Selecting features that intersect

The intersect operator identifies features that are not disjoint in the sense of Figure 8.13, but now extended to include points and lines. Figure 9.6b provides an example of spatial selection using the intersect relationship between lines and polygons. All roads intersecting Ilala District were selected.

Selecting features adjacent to selection objects

Adjacency is the meet relationship described in the subsection Topology and spatial relationships of Chapter 8 (page 249). The adjacency operator identifies those features that share boundaries and, therefore, applies only to line and polygon features. Figure 9.6c illustrates a spatial adjacency query. We want to select all parcels adjacent to an industrial area. The first step is to select the industrial area (in dark blue) and then
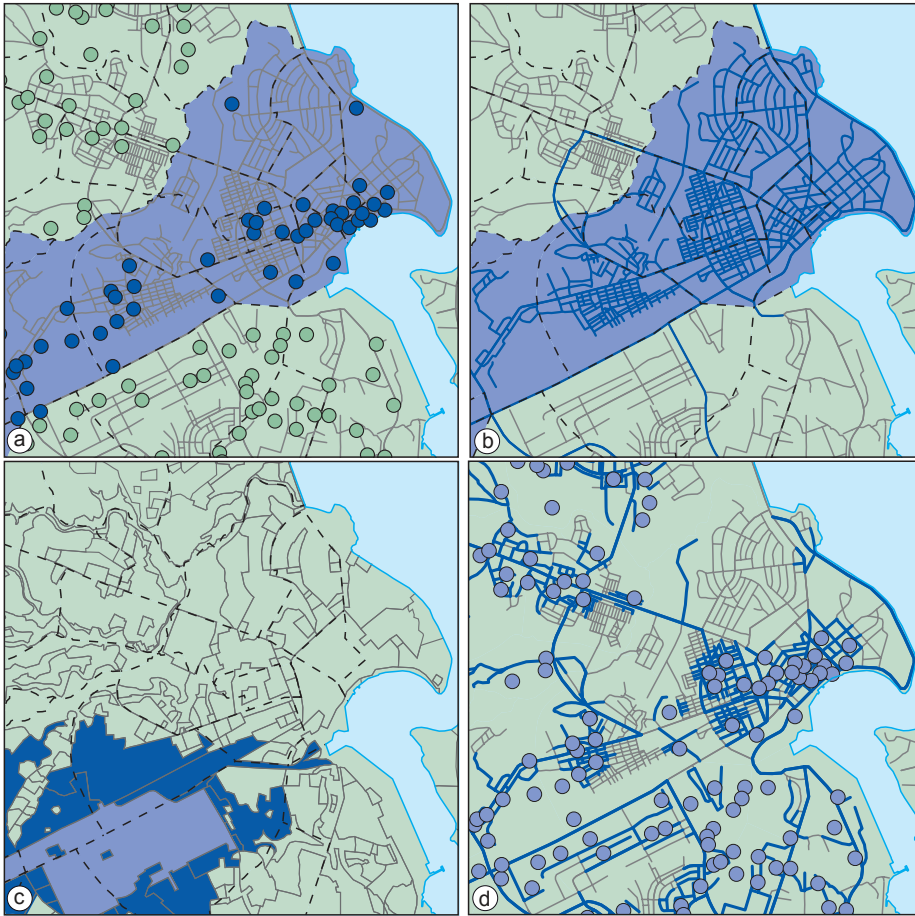
Figure 9.6
Spatial selection using topological relationships: (a) selection of medical clinics (dark purple) located inside the selected district (violet); (b) selection of roads (dark purple) that are partially located in the selected district (violet); (c) selection of the areas (dark purple), adjacent to the selected urban industrial area (violet); (d) selection of roads within 200 m of a clinic.

apply the adjacency function to select all land use areas (delineated in light blue) that are adjacent to it.

Selecting features based on their distance

One may also want to use the distance function of a GIS as a tool for selecting features. Such selections can be searches for features within a given distance of the selection objects, at a given distance, or even beyond a given distance. There is a whole range of selections of this type. For example:

- Which medical clinics are within 2 km of a selected school? (Information needed for the school's emergency procedures.)

- Which roads are within 200 m of a medical clinic? (These roads must have a high priority for road maintenance.)

Figure 9.6d illustrates a spatial selection using distance. In this case, we executed the selection of the second example directly above. Our selection objects were all clinics, and we selected the roads that pass within 200 m of a clinic. For situations in which we know the distance criteria to use—for selections within, at or beyond that distance value—the GIS has many (straightforward) computations to perform. Things

become more complicated if the distance selection condition involves the word "nearest" or "farthest". The reason is that not only must the GIS compute distances from a selection object A to all potentially selectable features F , but also that it must find the feature F that is nearest to (or farthest away from) object A. So, this requires an extra computational step to determine minimum (maximum) values. Most GIS packages support this type of selection, though the mechanics ("the buttons to use") differ among packages.

Afterthought on selecting features

So far we have discussed a number of different techniques for selecting features. We have also seen that selection conditions on attribute values can be combined using logical connectives such as AND, OR and NOT . Other techniques of selecting features can also usually be combined. Any set of selected features can be used as the input for a subsequent selection procedure. This means, for instance, that we can select all medical clinics first, then identify roads within 200 m of them, then select from those only the major roads, then select the nearest clinics to these remaining roads as the ones that should receive our financial support for maintenance. In this way, we are combining various techniques of selection.

9.2.3 Classification

Classification is a technique for purposely removing detail from an input data set in the hope of revealing important patterns (of spatial distribution). In the process, we produce an output data set, so that the input set can be left intact. This output set is produced by assigning a characteristic value to each element in the input set, which is usually a collection of spatial features that could be raster cells or points, lines or polygons. If the number of characteristic values in the output set is small in comparison to the size of the input set, we have classified the input set.

The pattern that we are looking for may be the distribution of household income in a city. In that case, household income is called the classification parameter. If we know for each ward in the city the associated average income, we have many different values. Subsequently, we could define five different categories (or classes) of income: "low", "below average", "average", "above average" and "high", and provide value ranges for each category. If these five categories are mapped using a sensible colour scheme, this may reveal interesting information. Figure 9.7 illustrates how this has been done in two ways for Dar es Salaam.

reclassification

The input data set may, itself, have been the result of a classification. In such cases we refer to the output data set as a reclassification. For example, we may have a soil map that shows different soil type units and we would like to show the suitability of units for a specific crop. In this case, it is better to assign to the soil units an attribute of suitability for the crop. Since different soil types may have the same crop suitability, a classification may merge soil units of different type into the same category of crop suitability.

aggregation and merging
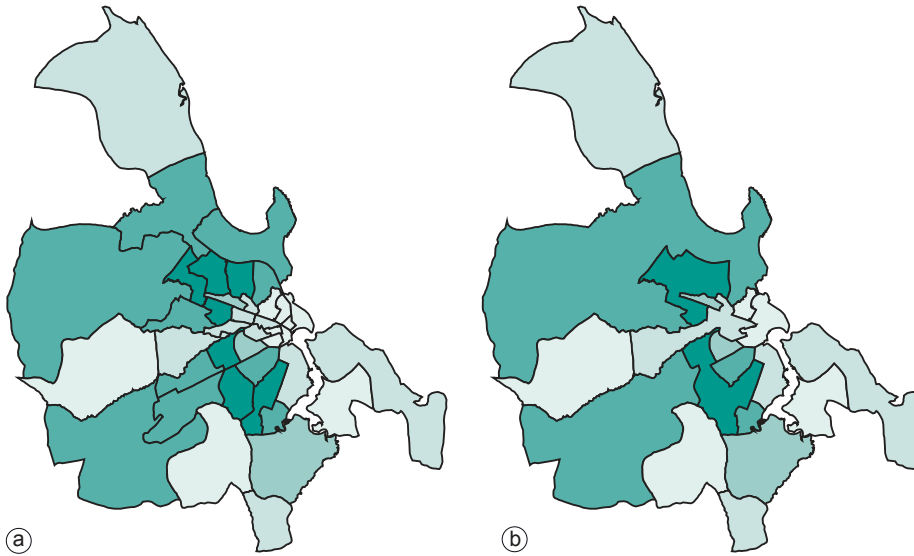
In classification of vector data, there are two possible results. In the first, the input features may become the output features in a new data layer, with an additional category assigned. In other words, nothing changes with respect to the spatial extents of the original features. Figure 9.7a illustrates this first type of output. A second type of output is obtained when adjacent features of the same category are merged into one bigger feature. Such a post-processing function is called spatial merging, aggregation or dissolving. An illustration of this second type is found in Figure 9.7b. Observe that this type of merging is only an option in vector data, as merging cells in an output raster on the basis of a classification makes little sense. Vector data classification can

be performed on point sets, line sets or polygon sets; the optional merge phase only
makes sense for lines and polygons. In the following two subsections we discuss two
kinds of classification: user-controlled and automatic classification.

Household income range	New category value
391–2474	1
2475–6030	2
6031–8164	3
8165–11587	4
11588–21036	5

User-controlled classification

In user-controlled classification, a user selects the attribute(s) that will be used as the
classification parameter(s) and defines the classification method. The latter involves
declaring the number of classes, as well as the correspondence between the old at-
tribute values and the new classes. This is usually done via a classification table. The
classification table used for Figure 9.7 is displayed in Table 9.1. It is rather typical
for cases in which the parameter domain used is continuous (e.g. household income).
Then, the table indicates value ranges to be mapped to the same category. Note that
category values are ordinal data, described in the subsection Geographic fields on
(page 239).

classification table

Another case exists when the classification parameter is nominal or at least discrete.
Such an example is given in Figure 9.8. We must also define the data format of the
output as a spatial data layer, which will contain the new classification attribute. The
data type of this attribute is always categorical, i.e. integer or string, no matter what
the data type of the attribute(s) from which the classification was obtained.

Sometimes, one may want to classify only a selection of features. In such cases, there
are two options for the features that are not selected. One option is to keep their
original values, while the other is to assign a null value to them in the output data
set. A null value is a special value that means that no applicable value is present.
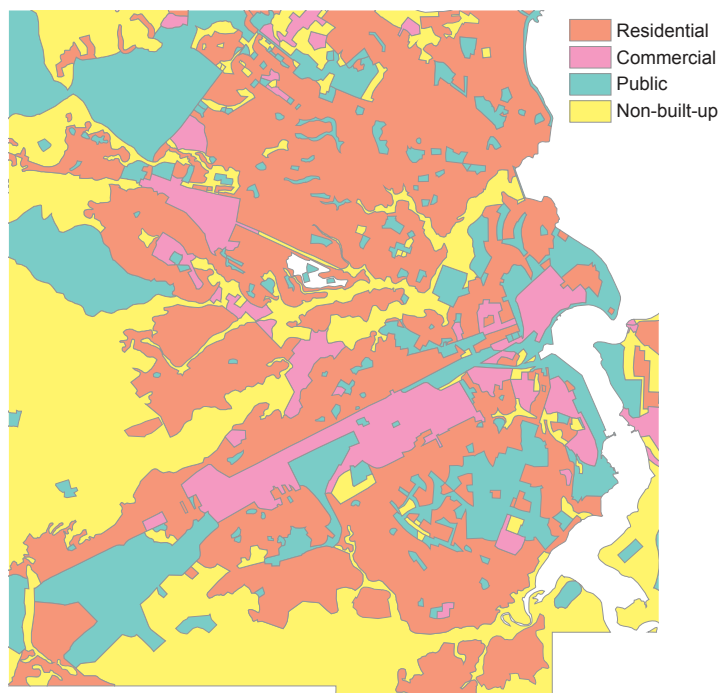
Figure 9.8
An example of a classification on a discrete parameter, namely land use unit in city of Dar es Salaam, Tanzania.

Care must be taken to deal with these values correctly, both in computations and in visualization.

Automatic classification

User-controlled classifications require a classification table or user interaction. GIS software can also perform automatic classification, in which a user only specifies the number of classes in the output data set. The system automatically determines the class break points. The two main techniques of determining break points being used are the equal interval technique and the equal frequency technique.

Equal interval technique

The minimum and maximum values v_{min} and v_{max} of the classification parameter are determined and the (constant) interval size for each category is calculated as $(v_{max} - v_{min})/n$, where n is the number of classes chosen by the user. This classification is useful in that it reveals the distribution pattern, as it determines the number of features in each category.

Equal frequency technique

This technique is also known as quantile classification. The objective is to create categories with roughly equal numbers of features per category. The total number of features is determined first, then, based on the required number of categories, the number of features per category is calculated. The class break points are then determined by counting off the features in order of classification parameter value.

Both techniques are illustrated on a small 5×5 raster in Figure 9.9.

(a) original raster (b) equal interval classification (c) equal frequency classification

original value	new value	# cells
1,2	1	9
3,4	2	8
5,6	3	3
7,8	4	3
9,10	5	2

original value	new value	# cells
1	1	6
2,3	2	5
4	3	6
5,6	4	3
8,9,10	5	5
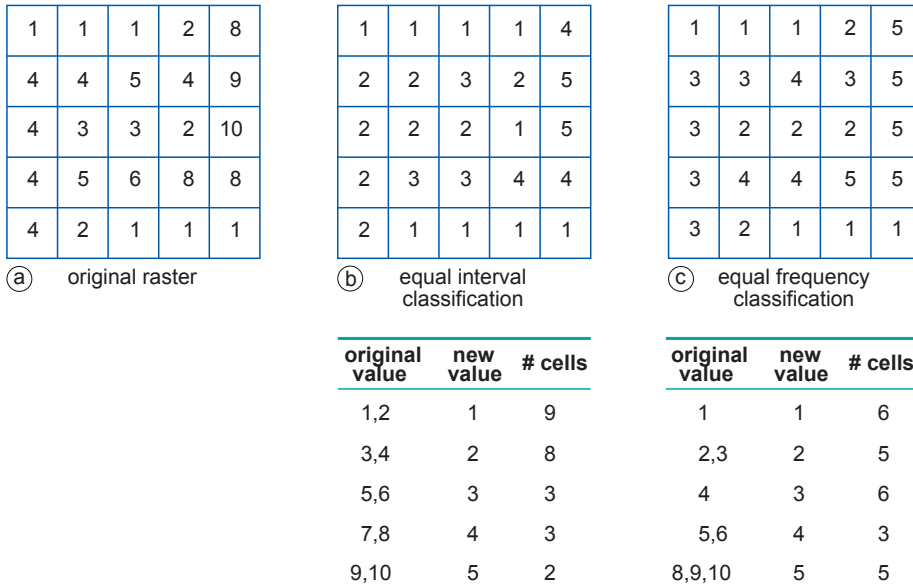
Figure 9.9
Example of two automatic classification techniques: (a) the original raster with cell values; (b) classification based on equal intervals; (c) classification based on equal frequencies. Below, the respective classification tables, with a tally of of the number of cells involved.

9.3 Overlay functions

In the previous section, various techniques for measuring and selecting spatial data were discussed. We also discussed the generation of a new spatial data layer from an old layer using classification. Now, in this section, we look at techniques for combining two spatial data layers and producing a third layer from them. The binary operators that we discuss are known as spatial overlay operators. Vector overlay operators will be dealt with first, followed by raster operators.

Standard overlay operators take two input data layers and assume that they are geo-referenced in the same system and that they overlap in the study area. If either of these requirements is not met, the use of an overlay operator is pointless. The principle of spatial overlay is to compare the characteristics of the same location in both data layers and to produce a result for each location in the output data layer. The specific result to produce is determined by the user. It might involve a calculation or some other logical function to be applied to every area or location. With raster data, as we shall see, these comparisons are carried out between pairs of cells, one from each input raster. With vector data, the same principle of comparing locations applies but the underlying computations rely on determining the spatial intersections of features from each input layer.

9.3.1 Vector overlay operators

In the vector domain, overlay is computationally more demanding than in the raster domain. Here we will only discuss overlays from polygon data layers, but do note that most of the ideas also apply to overlay operations with point or line data layers. The standard overlay operator for two layers of polygons is the polygon intersection operator. It is fundamental, as many other overlay operators proposed in the literature or implemented in systems can be defined in terms of it. The principles are illustrated in Figure 9.10. The result of this operator is the collection of all possible polygon intersections; the attribute table result is a join—in the relational database sense of Section 8.4—of the two input attribute tables. This output attribute table only con-

intersect

tains one tuple for each intersecting polygon found, which explains why we call this operator a spatial join. A more practical example is provided in Figure 9.11a, which was produced by polygon intersection of the ward polygons with land use polygons classified as in Figure 9.8. This allows us to select the residential areas in Ilala District.
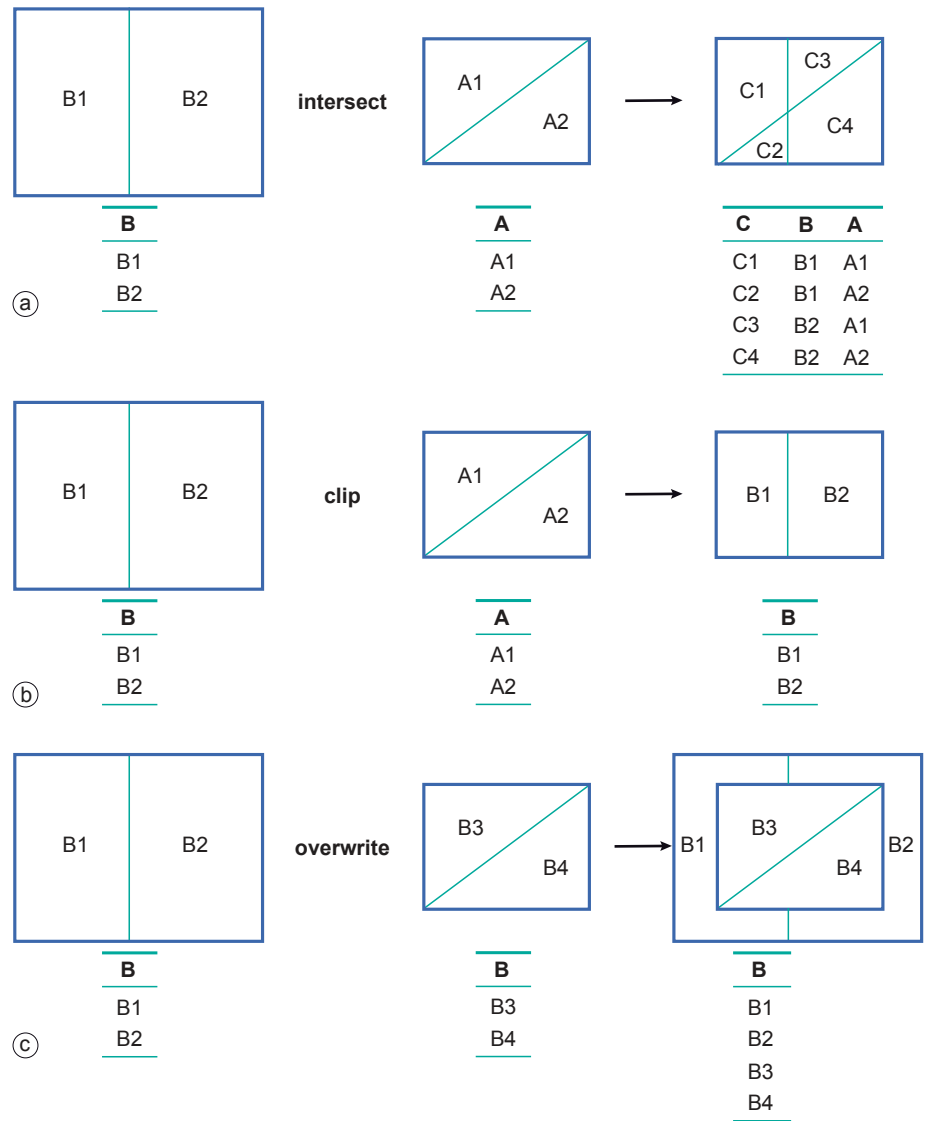
Figure 9.10
The polygon overlay operators: (a) polygon intersection, (b) polygon clipping, the left hand polygon (B) is clipped by polygon A, (c) polygon overwrite.

Three polygon overlay operators are illustrated in Figure 9.10. The second is known as the polygon clipping operator. It takes a polygon data layer and restricts its spatial extent to the generalized outer boundary obtained from all (selected) polygons in a second input layer. Besides this generalized outer boundary, no other polygon boundaries from the second layer play a role in the result.

The third overlay operator illustrated in Figure 9.10 is referred to as polygon overwrite. The result of this binary operator is a polygon layer with the polygons of the first layer, except where polygons exist in the second layer, as these take priority.

clip

overwrite

Most GISs do not force the user to apply overlay operators to the full polygon data set. One is allowed to first select relevant polygons in the data layer and then use the selected set as an operator argument. The fundamental operator of all these is polygon intersection. The other operators can be defined in terms of it, usually in combination with polygon selection and/or classification. For instance, the polygon overwrite of A by B can be defined as polygon intersection between A and B, followed by a (well-chosen) classification that prioritizes polygons in B, followed by a merge. (The reader is asked to verify this.) Vector overlays are usually also defined for point or line data layers. Their definition parallels the definitions of operators discussed above. Different GISs use different names for these operators, so it is advisable to carefully check the documentation before applying any of these operators.

9.3.2   Raster overlay operators

Vector overlay operators are useful but geometrically complicated, and this sometimes results in poor operator performance. Raster overlays do not suffer from this disadvantage, as most of them perform their computations cell by cell, and thus they are fast. GISs that support raster processing—as most do—usually have a language to express operations on rasters. These languages are generally referred to as map algebra [111] or, sometimes, raster calculus. They allow a GIS to compute new rasters from existing ones, using a range of functions and operators. Unfortunately, not all implementations of map algebra offer the same functionality. The discussion below is to a large extent based on general terminology; it attempts to illustrate the key operations using a logical, structured language. Again, the syntax often varies among different GIS software packages.

map algebra

When producing a new raster we must provide a name for it, and define how it is to be computed. This is done in an assignment statement of the following format:

Output raster name := Map algebra expression.

The expression on the right is evaluated by the GIS, and the raster in which it results is then stored under the name on the left. The expression may contain references to existing rasters, operators and functions; the format is made clear in each case. The raster names and constants that are used in the expression are called its operands. When the expression is evaluated, the GIS will perform the calculation on a pixel-by-pixel basis, starting from the first pixel in the first row and continuing through to the last pixel in the last row. In map algebra, there is a wide range of operators and functions available, some of which will be discussed in the following subsections.

Arithmetic operators

Various arithmetic operators are supported. The standard ones are multiplication ($\times$), division (/), subtraction ($-$) and addition ($+$). Obviously, these arithmetic operators should only be used on appropriate data values, and, for instance, not on classification values. Other arithmetic operators may include modulo division (MOD) and integer division (DIV). Modulo division returns the remainder of division: for instance, 10 MOD 3 will return 1 as $10 - 3 \times 3 = 1$. Similarly, 10 DIV 3 will return 3.

Figure 9.12
Examples of arithmetic map algebra expressions.
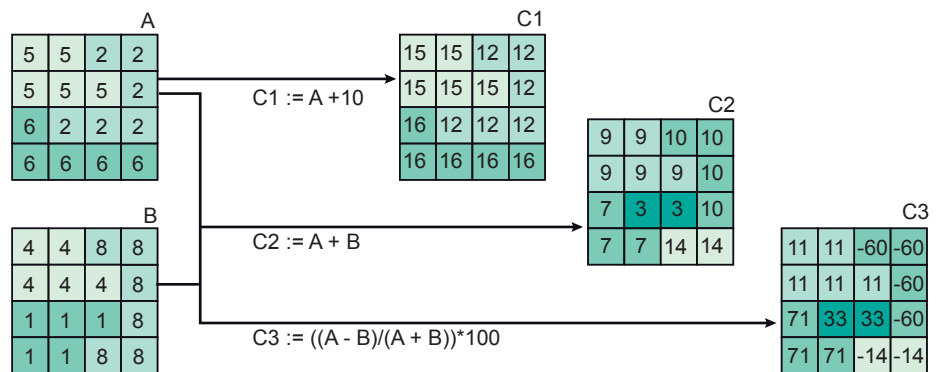
Other operators are goniometric: sine (sin), cosine (cos), tangent (tan); and their inverse functions asin, acos, and atan, which return radian angles as real values. Some simple map algebra assignments are illustrated in Figure 9.12. The assignment

$$C1 := A + 10$$

will add a constant factor of 10 to all cell values of raster A and store the result as output raster $C1$. The assignment

$$C2 := A + B$$

will add the values of A and B cell by cell, and store the result as raster $C2$. Finally, the assignment

$$C3 := (A - B)/(A + B) \times 100$$

will create output raster $C3$, as the result of the subtraction (cell by cell, as usual) of B cell values from A cell values, divided by their sum. The result is multiplied by 100. This expression, when carried out on AVHRR channel 1 (red) and AVHRR channel 2 (near infrared) of NOAA satellite imagery, is known as the NDVI (Normalized Difference Vegetation Index). It has proven to be a good indicator of the presence of green vegetation.
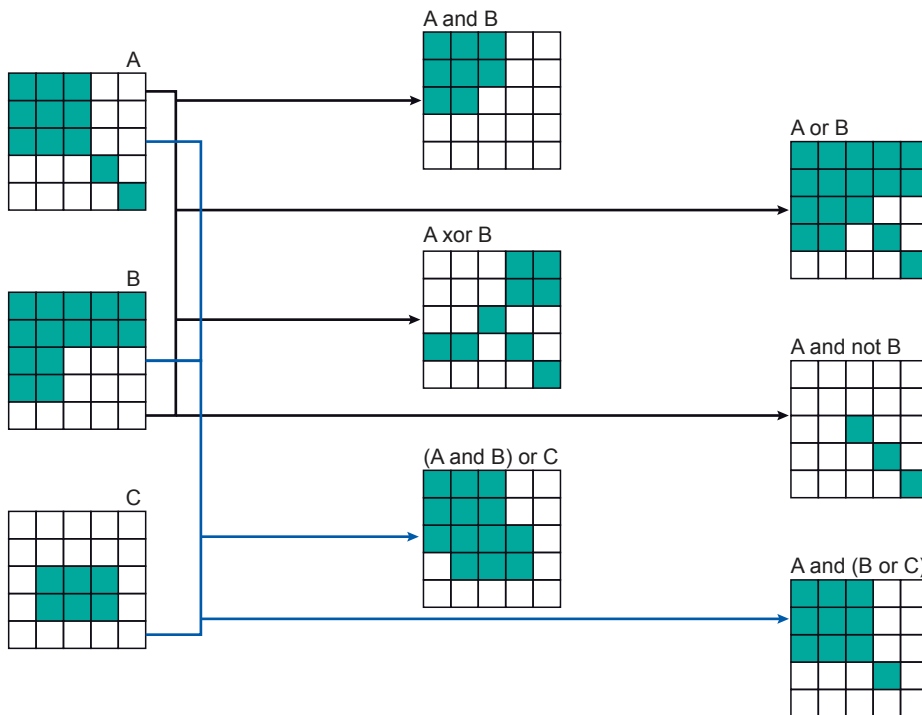
Figure 9.13
Example of logical expression in map algebra. Green cells represent true values, white cells represent false values.

Comparison and logical operators

Map algebra also allows the comparison of rasters cell by cell. To this end, we may use the standard comparison operators (<, <=, =, >=, > and <>) that were introduced in Subsection 9.3.2.

A simple raster comparison assignment is

$C := A <> B.$

It will store truth values—either true or false—in the output raster C. A cell value in C will be true if the cell's value in A differs from that cell's value in B. It will be false if they are the same. Logical connectives are also supported in many raster calculi. We have already seen the connectives of AND , OR and NOT in raster overlay operators (page 327). Another connective that is commonly offered in map algebra is exclusive OR (XOR). The expression a XOR b is true only if either a or b is true, but not both.

Examples of the use of these comparison operators and connectives are provided in Figure 9.13 and Figure 9.14. The latter figure provides various raster computations in searches for forests at specific elevations. In the figure, raster $D1$ indicates forest below 500 m, $D2$ indicates areas below 500 m or that are forests, raster $D3$ areas that are either forest or below 500 m (but not at the same time), and raster $D4$ indicates forests above 500 m.

Conditional expressions

The above comparison and logical operators produce rasters with the truth values true and false. In practice, we often need a conditional expression together with them that allows us to test whether a condition is fulfilled. The general format is:

Output raster := CON(condition, then expression, else expression).

Figure 9.14
Examples of complex logical expressions in map algebra. *A* is a classified raster for land use, and *B* holds elevation values.
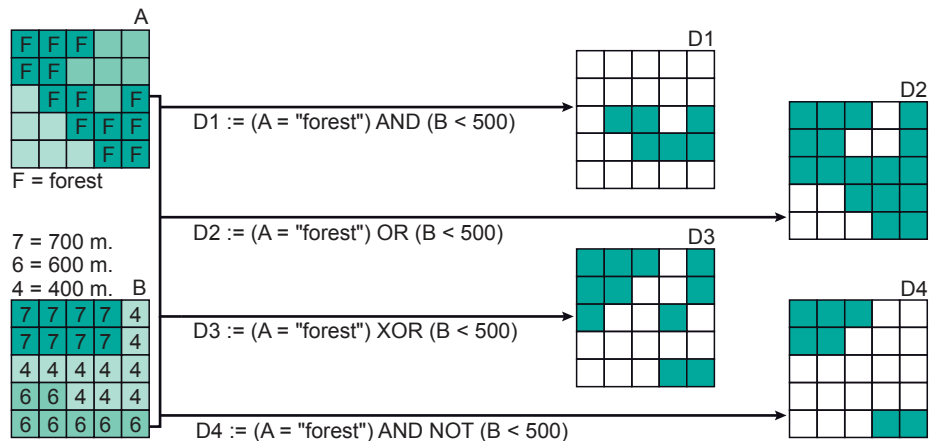
Here, condition stands for the condition tested, then the expression is evaluated if condition holds, and else the expression is evaluated if it does not hold. This means that an expression such as CON(A = "forest", 10, 0) will evaluate to 10 for each cell in the output raster where the same cell in A is classified as forest. For each cell where this is not true, the else expression is evaluated, resulting in 0. Another example is provided in Figure 9.15, showing that values for the *then* expression and the *else* expression can be some integer (possibly derived from another calculation) or values derived from other rasters. In this example, the output raster $C1$ is assigned the values of input raster B wherever the cells of input raster A contain forest. The cells in output raster $C2$ are assigned 10 wherever the elevation (B) is equal to 7 and the ground cover (A) is forest.

Figure 9.15
Examples of conditional expressions in map algebra. Here A is a classified raster holding land use data, and B is an elevation-value raster.
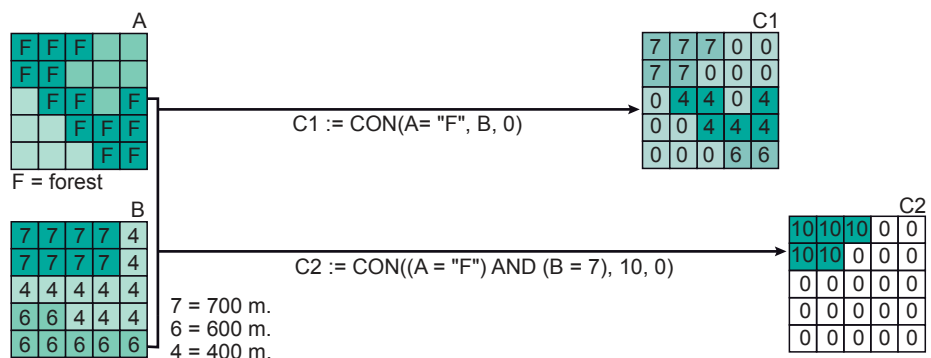
Overlays using a decision table

Conditional expressions are powerful tools in cases where multiple criteria must be taken into account. A small example may illustrate this. Consider a suitability study in which a land use classification and a geological classification must be used. The respective rasters are shown on the left-hand side of Figure 9.16. Domain expertise dictates that some combinations of land use and geology result in suitable areas, whereas other combinations do not. In our example, forests on alluvial terrain and grassland on shale are considered suitable combinations, while any others are not.

We could produce the output raster of Figure 9.16 with a longish map algebra expression, such as

Suitability := CON((Landuse = "Forest" AND Geology = "Alluvial")
OR (Landuse = "Grass" AND Geology = "Shale"),
"Suitable", "Unsuitable")

and consider ourselves lucky that there are only two "suitable" cases. In practice, many more cases must usually be covered and, then, writing up a complex CON expression is not an easy task.

To this end, some GISs accommodate setting up a separate decision table that will guide the raster overlay process. This extra table carries domain expertise and dictates which combinations of input raster-cell values will produce which output raster-cell value. This gives us a raster overlay operator using a decision table, as illustrated in Figure 9.16. The GIS will have supporting functions to generate the additional table from the input rasters and to enter appropriate values in the table.
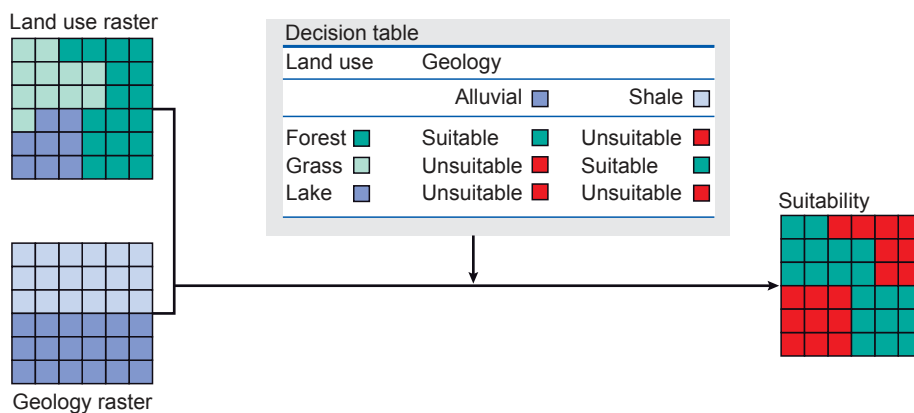
Figure 9.16
The use of a decision table in a raster overlay. The overlay would be computed in a suitability study in which land use and geology are important factors. The meaning of values in both input rasters, as well as the output raster, can be derived from the decision table.

9.4 Neighbourhood functions

In the explanation of overlay operators (Subsection 9.3.2), the guiding principle was to compare or combine the characteristic value of a location from two data layers and to do so for all locations. This is what map algebra, for instance, gives us: cell by cell calculations with the results stored in a new raster.

There is another guiding principle in spatial analysis that can be equally useful. The principle in this case is to find out the characteristics of the vicinity, here called neighbourhood, of a location. After all, many suitability questions, for instance, depend not only on what is at a location but also on what is near the location. Thus, the GIS must allow us "to look around locally". To perform neighbourhood analysis, we must:

1. state which target locations are of interest to us and define their spatial extent;

2. define how to determine the neighbourhood for each target; and

3. define which characteristic(s) must be computed for each neighbourhood.

For instance, our target might be a medical clinic. Its neighbourhood could be defined as:

- an area within a radius of 2 km distance as the crow flies; or

- an area within 2 km travelling distance; or

- all roads within 500 m travelling distance; or

- all other clinics within 10 minutes travelling time;

- all residential areas for which the clinic is the closest clinic.

Finally, in the third step we indicate what it is we want to discover about the phenomena that exist or occur in the neighbourhood. This might simply be its spatial extent, but it might also be statistical information such as:

- how many people live in the area;

- what is their average household income;

- are any high-risk industries located in the neighbourhood.

These are typical questions in an urban setting. When our interest is more in natural phenomena, different examples of locations, neighbourhoods and neighbourhood characteristics arise. Since raster data are the more commonly used in this case, neighbourhood characteristics often are obtained via statistical summary functions that compute values such as the average, minimum, maximum and standard deviation of the cells in the identified neighbourhood.

To select target locations, one can use the selection techniques that we discussed in Section 9.3. To obtain characteristics from an eventually-to-be identified neighbourhood, the same techniques apply. So what remains to be discussed here is the proper determination of a neighbourhood. One way of determining a neighbourhood around a target location is by making use of the geometric distance function. Some of these techniques are discussed in Section 9.5. Geometric distance does not take into account direction, but certain phenomena can only be studied by doing so. Think of the spreading of pollution by rivers, groundwater flow or prevailing weather systems. The more advanced techniques for computation of flow and diffusion are discussed in Section 9.5.

geometric distance

Diffusion functions are based on the assumption that the phenomenon in question spreads in all directions, though not necessarily equally easily in each direction. Hence it uses local terrain characteristics to compute local resistances to diffusion. In flow computations, the assumption is that the phenomenon will choose a path of least-resistance and will not spread in all directions. This, as we will see, involves the computation of preferred local direction of spread. Both flow and diffusion computations take local characteristics into account and are, therefore, more easily performed on raster data.

9.4.1 Proximity computations

In proximity computations, we use geometric distance to define the neighbourhood of one or more target locations. The most common and useful technique is buffer zone generation. Another technique based on geometric distance that will also be discussed is Thiessen polygon generation.

Buffer zone generation

The principle of buffer zone generation is simple: we select one or more target locations and then determine the area around them within a certain distance. In Figure 9.17a, a number of main and minor roads were selected as targets and 75 m and 25 m (respectively) buffers were computed from them.

In some case studies, zoned buffers must be determined, for instance in assessments of the effects of traffic noise. Most GISs support this type of zoned-buffer computation. An illustration is provided in Figure 9.17b.
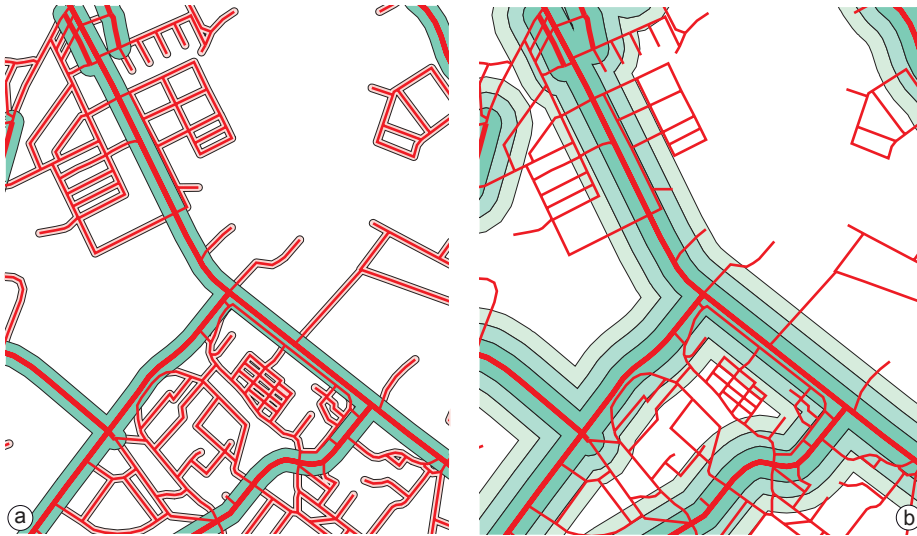
Figure 9.17
Buffer zone generation: (a) around main and minor zones. Different distances were applied: 25 m for minor roads, 75 m for main roads. (b) Zoned buffer zones around main roads. Three different zones were obtained: at 100 m, 200 m and 300 m from the main road.

In vector-based buffer generation, the buffers themselves become polygon features, usually in a separate data layer, that can be used in further spatial analysis. Buffer generation on rasters is a fairly simple function. The target location or locations are always represented by a selection of the raster's cells and geometric distance is defined using cell resolution as the unit. The distance function applied is the Pythagorean distance between the cell centres. The distance from a non-target cell to the target is the minimal distance one can find between that non-target cell and any target cell.

Thiessen polygon generation

Thiessen polygon partitions make use of geometric distance to determine neighbourhoods. This is useful if we have a spatially distributed set of points as target locations and we want to know the closest target for each location in the study. This technique will generate a polygon around each target location that identifies all those locations that "belong to" that target. We have already seen the use of Thiessen polygons in the context of interpolation of point data. Given an input point set that will be the polygon's midpoints, it is not difficult to construct such a partition. It is even much easier to construct if we already have a Delaunay triangulation for the same input point set (see page 246).

Delaunay triangulation   Thiessen polygons
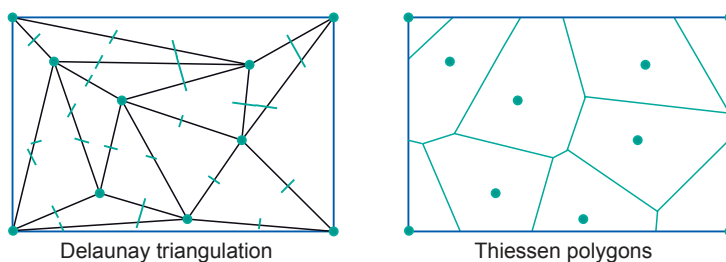
Figure 9.18
Thiessen polygon construction (right) from a Delaunay triangulation; (left): perpendiculars of the triangles form the boundaries of the polygons.

Figure 9.18 repeats the Delaunay triangulation of Figure 8.7b; the Thiessen polygon

partition constructed from it is on the right. The construction first creates the perpendiculars of all the triangle sides; note that a perpendicular of the side of a triangle that connects point A with point B is the imaginary line dividing the area between the area closer to A and the area closer to B. The perpendiculars become part of the boundary of each Thiessen polygon computed by the GIS. (The GIS will work for higher-precision real arithmetic rather than for what is illustrated here.)

9.4.2 Computation of diffusion

The determination of the neighbourhood of one or more target locations may depend not only on distance—cases of which we have discussed above—but also on direction and differences in the terrain in different directions. This is typically the case when the target location contains "source material" that spreads over time, referred to as *diffusion*. This "source material" may be air, water, soil pollution, commuters exiting a train station, people from an refugee camp that has just been opened up , a natural spring on a hillside , or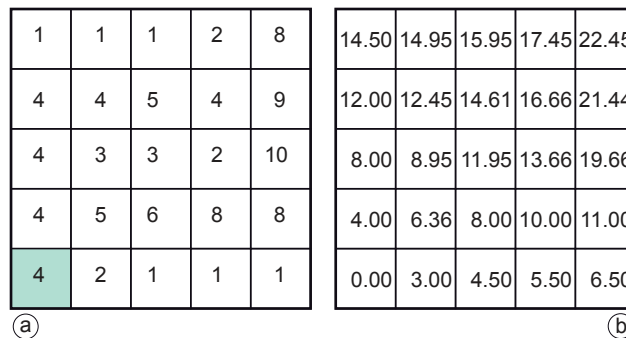 radio waves emitted from a radio relay station. In all these cases, one will not expect the spread to occur evenly in all directions. There will be local factors that influence the spread, making it easier or more difficult. Many GISs provide support for this type of computation. We will discuss some of the principles here—in the context of raster data.

diffusion and spread

Diffusion computation involves one or more target locations, which in this context are better called source locations: they are the locations of the source of whatever spreads. The computation also involves a local resistance raster, which for each cell provides a value that indicates how difficult it is for the "source material" to pass through that cell. The value in the cell must be normalized, i.e. valid for a standardized length (usually the cell's width) of spread path. From the source location(s) and the local resistance raster, the GIS will be able to compute a new raster that indicates how much minimal total resistance the spread has undergone before reaching a raster cell. This process is illustrated in Figure 9.19.

resistance raster

Figure 9.19
Computation of diffusion on a raster. The lower-left green cell is the source location, indicated in the local resistance raster (a). The raster in (b) is the minimal total resistance raster computed by the GIS. (The GIS will work in real arithmetic of higher precision than that illustrated here.)

1	1	1	2	8
4	4	5	4	9
4	3	3	2	10
4	5	6	8	8
4	2	1	1	1

ⓐ

14.50	14.95	15.95	17.45	22.45
12.00	12.45	14.61	16.66	21.44
8.00	8.95	11.95	13.66	19.66
4.00	6.36	8.00	10.00	11.00
0.00	3.00	4.50	5.50	6.50

ⓑ

While computing total resistances, the GIS takes proper care of the path lengths. Obviously, the diffusion from a cell c_{src} to its neighbour cell to the east c_e is shorter than to the cell that is its northeastern neighbour c_{ne}. The distance ratio between these two cases is $1/\sqrt{2}$. If $val(c)$ indicates the local resistance value for cell c, the GIS computes the total incurred resistance for diffusion from c_{src} to c_e as $1(val(c_{src}) + val(c_e))$, while the same for c_{src} to c_{ne} is $1(val(csrc) + val(cne)) \times \sqrt{2}$. The accumulated resistance along a path of cells is simply the sum of these incurred resistances from pairwise neighbour cells.

Since "source material" has the habit of taking the easiest route to spread, we must determine at what minimal "cost" (i.e. at what minimal resistance) it may have arrived

in a cell. Therefore, we are interested in the minimal cost path. To determine the minimal total resistance along a path from the source location c_{src} to an arbitrary cell c_x, the GIS determines all possible paths from c_{src} to c_x and then determines which one has the lowest total resistance. This value is found, for each cell, in the raster of Figure 9.19b.

For instance, there are three paths from the green source location to its northeast neighbour cell (with local resistance 5). We can define them as path 1 (N–E), path 2 (E–N) and path 3 (NE), using compass directions to define the path from the green cell. For path 1, the total resistance is computed as:

$$1/2(4 + 4) + 1/2(4 + 5) = 8.5.$$

Path 2, in similar style, gives us a total value of 6.5. For path 3, we find

$$1/2(4 + 5) \times \sqrt{2} = 6.36,$$

and, thus, it is obviously the minimal cost path. The reader is asked to verify one or two other values of minimal cost paths that the GIS has produced for the output raster.

9.4.3 Flow computation

Spread computations determine how a phenomenon spreads over an area—in principle in all directions, though with varying difficulty or resistance. There are also cases for which a phenomenon does not spread in all directions, but moves or "flows" along a given, least-cost path, determined again by characteristics of local terrain. The typical case arises when we want to determine drainage patterns in a catchment area: rain water "chooses" a way to leave the area.
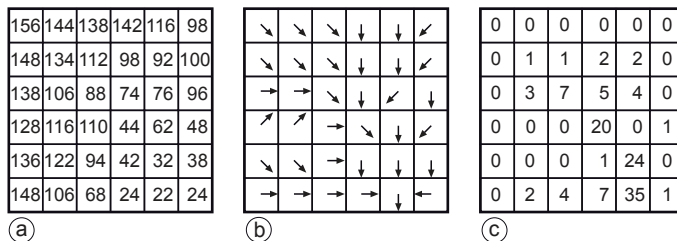
Figure 9.20
Flow computations on a raster: (a) the original elevation raster, (b) the flow-direction raster computed from it, (c) accumulated flow-count raster.

We can illustrate the principles involved in this typical case with the simple elevation raster provided in Figure 9.20a. For each cell in that raster, the steepest downward slope to a neighbour cell is computed and its direction is stored in a new raster (Figure 9.20b). This computation determines the elevation difference between the cell and the neighbour cell and it takes into account cell distance—1 for neighbour cells in N–S or W–E direction, 2 for cells in a NE–SW or NW–SE direction. From among its eight neighbour cells, it picks the one with the steepest path to it. The directions thus obtained in raster (b) are encoded in integer values; we have "decoded" them for the sake of illustration. Raster (b) can be called the flow-direction raster. From raster (b), the GIS can compute the accumulated flow- count raster, a raster that for each cell indicates how many cells have their water flow into that cell.

Cells with a high accumulated flow count represent areas of concentrated flow and may, thus, belong to a stream. By using some appropriately chosen threshold value in a map algebra expression, we may decide whether they do or not. Cells with an accumulated flow count of zero are local topographic highs and can be used to identify ridges.

9.4.4 Raster-based surface analysis

Continuous fields have a number of characteristics not shared by discrete fields. Since the field changes continuously, we can talk of slope angle, slope aspect and concavity/convexity of the slope. These notions are not applicable to discrete fields. The discussions in this subsection use terrain elevation as the prototype example of a continuous field, but all aspects discussed are equally applicable to other types of continuous fields. Nonetheless, we regularly refer to the continuous field representation as a DEM, to conform with the most common situation. Throughout the rest of this subsection we will assume that the DEM is represented as a raster.

Applications

There are numerous examples that require more advanced computations on continuous field representations, such as:

- Slope angle calculation—the calculation of the slope steepness, expressed as an angle in degrees or percentages, for any or all locations.

- Slope aspect calculation—the calculation of the aspect (or orientation) of the slope in degrees (between 0 and 360°), for any or all locations.

- Slope convexity/concavity calculation—defined as the change of the slope (negative when the slope is concave and positive when the slope is convex)—can be calculated as the second derivative of the field.

- Slope length calculation—with the use of neighbourhood operations, it is possible to calculate for each cell the nearest distance to a watershed boundary (the upslope length) and to the nearest stream (the downslope length). This information is useful for hydrological modelling.

- Hillshading is used to portray relief difference and terrain morphology of hilly and mountainous areas. The application of a special filter to a DEM produces hillshading. The colour tones in a hillshading raster represent the amount of reflected light at each location, depending on its orientation relative to the illumination source. This illumination source is usually chosen to be to the northwest at an angle of 45° above the horizon.

- Three-dimensional map display—with GIS software, three-dimensional views of a DEM can be constructed in which the location of the viewer, the angle under which he or she is looking, the zoom angle, and the amplification factor of relief exaggeration can be specified. Three-dimensional views can be constructed using only a predefined mesh, covering the surface, or using other rasters (e.g. a hillshading raster) or images (e.g. satellite images) that are draped over the DEM.

- Determination of change in elevation through time—the cut-and-fill volume of soil to be removed or to be brought in to make a site ready for construction can be computed by overlaying the DEM of the site before the work begins with the DEM of the expected modified topography. It is also possible to determine landslide effects by comparing DEMs of before and after a landslide event.

- Automatic catchment delineation—catchment boundaries or drainage lines can be automatically generated from a good quality DEM with the use of neighbourhood functions. The system will determine the lowest point in the DEM, which is considered to be the outlet of the catchment. From there, it will repeatedly

search for the neighbouring pixels with the highest altitude. This process is repeated until the highest location (i.e. the cell with the highest value) is found; the path followed determines the catchment boundary. For delineating the drainage network, the process is reversed. Then the system will work from the watershed downwards, each time looking for the lowest neighbouring cells, which determines the direction of water flow.

- Dynamic modelling—apart from the applications mentioned above, DEMs are increasingly used in GIS-based dynamic modelling, such as the computation of surface run-off and erosion, groundwater flow, the delineation of areas affected by pollution, the computation of areas that will be covered by processes such as flows of debris and lava.

- Visibility analysis—a viewshed is the area that can be "seen" (i.e. it is in the direct line-of-sight) from a specified target location. Visibility analysis can determine the area visible from a scenic lookout or the area that can be reached by a radar antenna, as well as assess how effectively a road or quarry will be hidden from view.

Some of the more important computations mentioned in the list above are discussed further in the following subsections. All of them apply a technique known as filtering, which has been described in Subsection 5.1.7.

Computation of slope angle and slope aspect

A different choice of weight factors may provide other information. Special filters exist to perform computations on the slope of the terrain. Before we look at these filters, let us define various notions of slope.
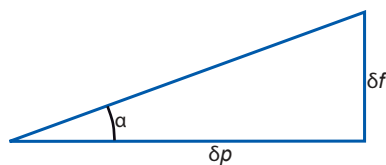
Figure 9.21
Slope angle defined. Here, δp stands for length in the horizontal plane, δf stands for the change in field value, where the field is usually terrain elevation. The slope angle is α.

Slope angle, also known as slope gradient, is the angle α illustrated in Figure 9.21, between a path p in the horizontal plane and the sloping terrain. The path p must be chosen such that the angle α is maximal. A slope angle can be expressed as elevation gain in a percentage or as a geometric angle, in degrees or radians. The two respective formulas are:

$$slope\_perc = 100\frac{\delta f}{\delta p} \text{ and } slope\_angle = \arctan\left(\frac{\delta f}{\delta p}\right)$$

The path p must be chosen to provide the highest slope-angle value and thus it can lie in any direction. The compass direction, converted to an angle to North, of this maximal downslope path p is what is called the slope aspect.

Let us now look at how to compute slope angle and slope aspect in a raster environment.

From an elevation raster, we cannot "read" the slope angle or slope aspect directly. Yet that information can be extracted. After all, for an arbitrary cell, we have its elevation value, plus those of its eight neighbouring cells. A simple approach to slope

angle computation is to make use of x- and y-gradient filters (gradient filters were introduced in Subsection 5.1.7). The x-gradient filter determines the slope increase ratio from west to east: if the elevation to the west of the centre cell is 1540 m and that to the east of the centre cell is 1552 m, then apparently along this transect the elevation increases 12 m per two cell widths, i.e. the x-gradient is 6 m per cell width. The y-gradient filter operates entirely analogously, though in the south-north direction.

Observe that both filters express elevation gain per cell width. This means that we must divide by the cell width—given in metres, for example—to obtain the (approximations to) the true derivatives $\delta f/\delta x$ and $\delta f/\delta y$. Here, f stands for the elevation field as a function of x and y, and $\delta f/\delta x$, for instance, is the elevation gain per unit of length in the x-direction.

To obtain the real slope angle α along path p, observe that both the x- and y-gradient contribute to it. This is illustrated in Figure 9.22. A not-so-simple geometric derivation can show that always:

$$\tan(\alpha) = \sqrt{\left(\frac{\delta f}{\delta x}\right)^2 + \left(\frac{\delta f}{\delta y}\right)^2}.$$

In the practice of computing local slope angles from an elevation raster, this means that we must perform the following steps:

1. Compute from (input) elevation raster R the non-normalized x- and y- gradients.

2. Normalize the resulting rasters by dividing by the cell width, expressed in units of length such as metres.

3. Use both rasters for generating a third raster, applying the $\sqrt{\ldots}$ formula above, and possibly even an `arctan` function to the result, to obtain the slope angle α for each cell.

It can also be shown that for the slope aspect ψ we have

$$\tan \psi = \frac{\delta f/\delta x}{\delta f/\delta y},$$
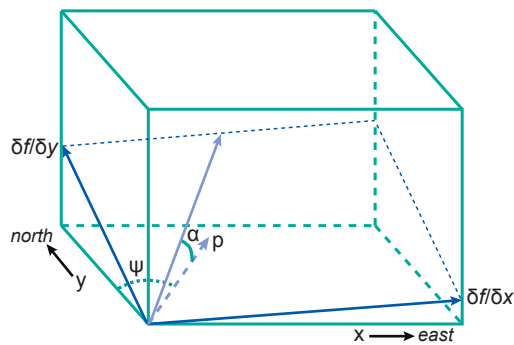
Figure 9.22
Slope angle and slope aspect defined. Here, p is the horizontal path in maximal slope direction and α is the slope angle. The plane tangent to the terrain in the origin is also indicated. The angle ψ is the slope aspect.

Slope aspect can, therefore, also be computed from the normalized gradients. Readers are warned that this formula should not be carelessly replaced by using

$$\psi = \arctan\left(\frac{\delta f/\delta x}{\delta f/\delta y}\right),$$

the reason being that the latter formula does not account for southeast and southwest quadrants, nor for cases where $\delta f/\delta y = 0$. (In the first situation, one must add 180° to the computed angle to obtain an angle measured from North; in the latter situation, ψ equals either 90° or −90°, depending on the sign of $\delta f/\delta x$.)

9.5 Network analysis

Computations on networks comprise a different set of analytical functions in GISs. A network is a connected set of lines representing some geographic phenomenon, typically to do with transportation. The "goods" transported can be almost anything: people, cars and other vehicles along a road network, commercial goods along a logistic network, phone calls along a telephone network, or water pollution along a stream/river network.

Network analysis can be performed on either raster or vector data layers, but they are more commonly done on the latter, as line features can be associated with a network and hence can be assigned typical transportation characteristics, such as capacity and cost per unit. A fundamental characteristic of any network is whether the network lines are considered to be directed or not. Directed networks associate with each line a direction of transportation; undirected networks do not. In the latter, the "goods" can be transported along a line in both directions. We discuss here vector network analysis, and assume that the network is a set of connected line features that intersect only at the lines' nodes, not at internal vertices. (But we do mention under- and overpasses.)

directed and undirected networks

For many applications of network analysis, a planar network, i.e. one that can be embedded in a two-dimensional plane, will do the job. Many networks are naturally planar, such as stream/river networks. A large-scale traffic network, on the other hand, is not planar: motorways have multi-level crossings and are constructed with underpasses and overpasses. Planar networks are easier to deal with computationally, as they have simpler topological rules. Not all GISs accommodate non-planar networks, or they can only do so using "tricks". These tricks may involve the splitting of overpassing lines at the intersection vertex and the creation of four lines from the two original lines. Without further attention, the network will then allow one to make a turn onto another line at this new intersection node, which in reality would be impossible. In some GISs we can allocate a cost for turning at a node—see our discussion on turning costs below—and that cost, in the case of the overpass trick, can be made infinite to ensure it is prohibited. But, as mentioned, this is a work around to fit a non-planar situation into a data layer that presumes planarity. The above is a good illustration of geometry not fully determining the network's behaviour. Additional application-specific rules are usually required to define what can and cannot happen in the network. Most GISs provide rule-based tools that allow the definition of these extra application rules. Various classical spatial analysis functions for networks are supported by GIS software packages. The most important ones are:

planar network—non-planar network

- optimal-path finding, which generates a least-cost path on a network between a pair of predefined locations using both geometric and attribute data.

- network partitioning, which assigns network elements (nodes or line segments) to different locations using predefined criteria.

These two typical functions are discussed in the two subsections that follow.
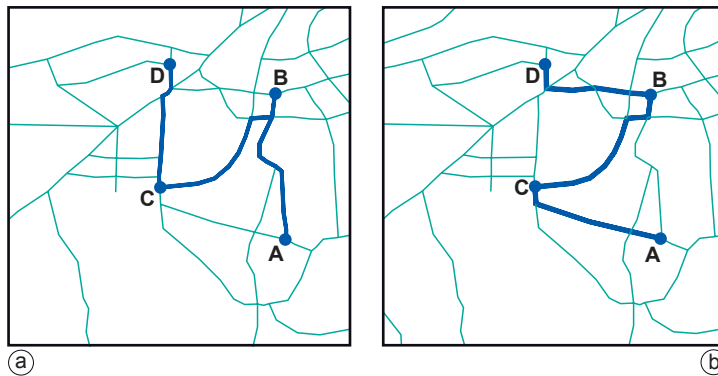
9.5.1 Optimal-path finding

Optimal-path finding techniques are used when a least-cost path between two nodes in a network must be found. The two nodes are called origin and destination. The aim is to find a sequence of connected lines to traverse from the origin to the destination at the lowest possible cost. The cost function can be simple: for instance, it can be defined as the total length of all lines of the path. The cost function can also be more elaborate and take into account not only length of the lines but also their capacity, maximum transmission (travel) rate and other line characteristics, for instance to obtain a reasonable approximation of travel time. There can even be cases in which the nodes visited add to the cost of the path as well. These may be called turning costs, which are defined in a separate turning-cost table for each node, indicating the cost of turning at the node when entering from one line and continuing on another. This is illustrated in Figure 9.23.

cost function

Figure 9.23
Network neighbourhood of node N with associated turning costs at N. Turning at N onto c is prohibited because of its direction, so no costs are mentioned for turning onto c. A turning cost of infinity (∞) also means that the turn is prohibited.

from	onto	cost
a	a	x
a	b	5
b	a	8
b	b	15
c	a	3
c	b	6
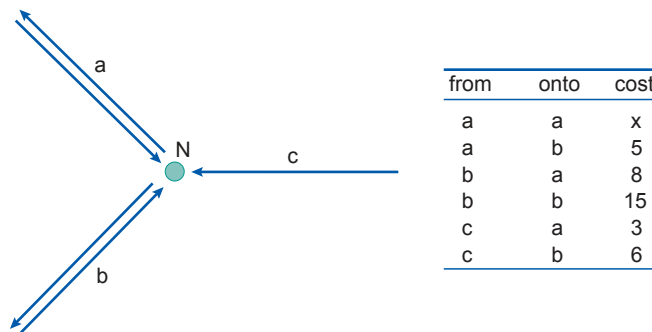
Figure 9.24
Ordered (a) and unordered (b) optimal-path finding. In both cases, a path had to be found from A to D: in (a) by visiting B and then C; in (b) also by visiting both nodes, but in arbitrary order.

The attentive reader will notice that it is possible to travel on line b in Figure 9.23, make a U-turn at node N, and return along a to where one came from. The question is whether doing this makes sense in optimal-path finding. After all, to go back to where one came from will only increase the total cost. In fact, there are situations where it is optimal to do so. Suppose it is node M that is connected by line b with node N, and that we actually wanted to travel from M to another node L. The turn at M towards node L coming via another line may be prohibitively expensive, whereas turning towards L at M and returning to M along b may not be so expensive.

ordered—unordered pathfinding

Problems related to optimal-path finding may require *ordered* optimal path finding or *unordered* optimal-path finding. Both have as an extra requirement that a number of additional nodes need to be visited along the path. In ordered optimal-path finding,

the sequence in which these extra nodes are visited matters; in unordered optimal-path finding it does not. An illustration of both types is provided in Figure 9.24. Here, a path is found from node A to node D, via nodes B and C. Obviously, the length of the path found under non-ordered requirements is at most as long as the one found under ordered requirements. Some GISs provide support for these more complicated path-finding problems.
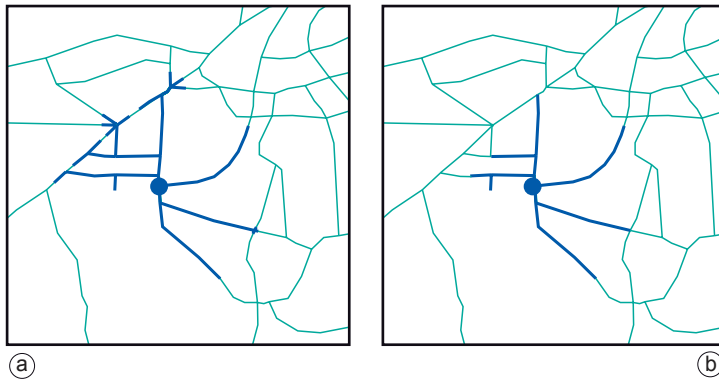
Figure 9.25
Network allocation for a pupil/school assignment problem. In (a), the street segments within 2 km of the school are identified; in (b), the selection of (a) is further restricted to accommodate the school's capacity for the new school year.

9.5.2 Network partitioning

In network partitioning, the purpose is to assign lines and/or nodes of the network in a mutually exclusive way to a number of target locations. Typically, the target locations play the role of service centres for the network. This may be any type of service, e.g. medical treatment, education, water supply. This sort of network partitioning is known as a network allocation problem. Another problem is *trace analysis*. Here, one wants to determine that part of a network that is upstream (or downstream) from a given target location. Such problems exist in tracing pollution along river/stream systems, but also in tracking down network failures in energy distribution networks.

9.5.3 Network allocation

In network allocation, we have a number of target locations that function as resource centres, and the problem is which part of the network to exclusively assign to which service centre. This may sound like a simple allocation problem, in which a service centre is assigned those line (segments) to which it is nearest, but usually the problem statement is more complicated. The additional complications stem from the requirements to take into account (a) the capacity with which a centre can produce the resources (whether they are medical operations, seats for school pupils, kilowatts or bottles of milk), and (b) the consumption of the resources, which may vary amongst lines or line segments. After all, some streets have more accidents, more children who live there, more industry in high demand of electricity or just more thirsty workers.

The service area of any centre is a subset of the distribution network, in fact a connected part of the network. Various techniques exist to assign network lines, or their segments, to a centre. In Figure 9.25a, the blue dot indicates a primary school and the GIS has been used to assign streets and street segments along the network within 2 km distance of the school. Then, using demographic figures on pupils living along the streets, it was determined that too many potential pupils lived in the area for the school's capacity. So in part (b), the part of the network already selected was reduced to match precisely the school's capacity for pupils in the new school year.

service area

9.5.4 Trace analysis

Trace analysis is performed when we want to understand which part of a network is "conditionally connected" to a chosen node on the network, which is known as the "trace origin". If a node or line is conditionally connected, this means that a path exists from the node/line to the trace origin, and that the connecting path fulfills the conditions set. What these conditions are depends on the application; they may involve the direction of the path, its capacity, its length, or resource consumption along it. The condition is typically a logical expression, as we have seen before:

- the path must be directed from the node/line to the trace origin;

- its capacity (defined as the minimum capacity of the lines that constitute the path) must be above a given threshold; and

- the path's length must not exceed a given maximum length.

Tracing is the computation that the GIS performs to find the paths from the trace origin that obey the tracing conditions. It is a rather useful function for many network-related problems. In Figure 9.26 the trace origin is indicated in red. In part (a), tracing conditions were set to trace all the way upstream; part (b) traces all the way downstream; and in part (c) there were no conditions set for the direction of the path, thereby tracing all connected lines from the trace origin. More complex conditions are certainly possible in tracing.
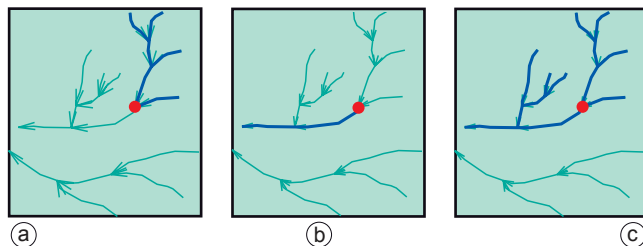
Figure 9.26
Tracing functions on a network: (a) tracing upstream, (b) tracing downstream, (c) tracing all connected lines from the origin.

9.6 Error propagation in spatial data processing

9.6.1 How errors propagate

In Section 8.5, a number of sources of error that may be present in source data were discussed. The acquisition of high quality base data still does not guarantee that the results of further, complex processing can be treated with certainty. As the number of processing steps increases, it becomes more difficult to predict the behaviour of such error propagation. These various errors may affect the outcome of spatial data manipulations. In addition, further errors may be introduced during the various processing steps discussed earlier in this chapter (see Figure 9.27).

Table 9.2 lists some common sources of error that may be introduced into GIS analyses. Note that these originate in a wide range of sources and include various common tasks relating to both data preparation and data analysis. It is the combination of different errors that are generated at each stage of preparation and analysis that may result in several errors and uncertainties in the final outputs.

One of the most commonly applied operations in GISs is analysis by overlaying two or more spatial data layers. Each of these layers will contain errors, due to both in-

Coordinate adjustments	*Generalization*
rubber sheeting/transformations	linear alignment
projection changes	line simplification
datum conversions	addition/deletion of vertices
rescaling	linear displacement
Feature Editing	*Raster/Vector Conversions*
line snapping	raster cells to polygons
extension of lines to intersection	polygons to raster cells
reshaping	assignment of point attributes
moving/copying	to raster cells
elimination of spurious polygons	post-scanner line thinning
Attribute editing	*Data input and Management*
numeric calculation and change	digitizing
text value changes/substitution	scanning
re-definition of attributes	topological construction / spatial indexing
attribute value update	dissolving polygons with same attributes
Boolean Operations	*Surface modelling*
polygon on polygon	contour/lattice generation
polygon on line	TIN formation
polygon on point	Draping of data sets
line on line	Cross-section/profile generation
overlay and erase/update	Slope/aspect determination
Display and Analysis	*Display and Analysis*
cluster analysis	class intervals choice
calculation of surface lengths	areal interpolation
shortest route/path computation	perimeter/area size/volume computation
buffer creation	distance computation
display and query	spatial statistics
adjacency/contiguity	label/text placement

herent inaccuracies in the source data and errors arising from some form of computer processing—for instance, rasterization. During the process of spatial overlay-

ing, errors in the individual data layers contribute to the final error of the output. The amount of error in the output depends on the type of overlay operation applied and on the amount of error in the individual layers. For example, errors in the results of an overlay using the logical operator AND are not the same as those created using the OR operator.

Consider another example. A land use planning agency is faced with the problem of identifying areas of agricultural land that are highly susceptible to erosion. Such areas occur on steep slopes in areas of high rainfall. The spatial data used in a GIS to obtain this information might include:

- a land use map produced five years previously from 1:25,000 scale aerial photographs;

- a DEM produced by interpolating contours from a 1:50,000 scale topographic map; and

- annual rainfall statistics collected with two rainfall gauges.

The reader is invited to assess what sort of errors are likely to occur in this analysis.

Referring back to Figure 9.27, the reader is also encouraged to reflect on errors introduced in the components of the application models discussed in the previous section, specifically, the methodological aspects of representing geographic phenomena. What might be the consequences of using a random function in an urban transportation model (when, in fact, travel behaviour is not purely random)?

9.6.2 Quantifying error propagation

Chrisman [18] noted that "the ultimate arbiter of cartographic error is the real world, not a mathematical formulation". We will never be able to capture and represent everything that happens in the real world perfectly in a GIS. Hence there is much to recommend the use of testing procedures for assessing accuracy. Various perspectives, motives and approaches for dealing with uncertainty have given rise to a wide range of conceptual models and indices for the description and measurement of error in spatial data. All these approaches have their origins in academic research and have solid theoretical foundations in mathematics and statistics. Here we identify two main approaches for assessing the nature and amount of error propagation:

1. testing the accuracy of each state by measurement against the real world; and

2. modelling error propagation, either analytically or by means of simulation techniques.

Error propagation can be modelled mathematically, although these models are very complex and only valid for certain types of data(e.g. numerical attributes). Rather than explicitly modelling error propagation, it is often more practical to test the results of each step in the process against some independently measured reference data.

Models of error and error propagation

It is important to distinguish models of error from models of error propagation in GISs. Modelling of error propagation has been defined by Veregin [118] as: "the application of formal mathematical models that describe the mechanisms whereby errors in source data layers are modified by particular data transformation operations."

In other words, we would like to know how errors in the source data behave under the manipulations that we subject them to in a GIS. If we are able to quantify the error in

the source data as well as their behaviour under GIS manipulations, we have a means of judging the uncertainty of their results.

Initially, error propagation models described only the propagation of attribute error [44], [118]. More recent research has addressed the spatial aspects of error propagation and the development of models incorporating both attribute and location components. These topics are beyond the scope of this book and readers are referred to [2] and [58] for more details.

Chapter 10

Visualization and dissemination

Rolf de By
Otto Huisman
Menno-Jan Kraak

10.1 Visualization

10.1.1 GISs and maps

There is a strong relationship between maps and GISs. Maps play a role at any moment one uses a GIS. They can be used as input, to verify data, to prepare a spatial analysis and of course to present results. As soon as a "where?" crops up in a question, a map can often be the most suitable tool for solving the question and providing the answer. "Where do I find Enschede?" and "Where do ITC's students come from?" are both examples. Of course, the answers could be in non-map form, such as "in the Netherlands" or "from all over the world". These answers could be satisfying enough. However, it is clear that these answers do not give a full picture. A map would put these answers in a spatial context. It could show where in the Netherlands Enschede is to be found and where it is located with respect to Schiphol–Amsterdam airport, where most students arrive when they come to the Netherlands. A world map would refine the answer "from all over the world," since it reveals that most students come from Africa and Asia, and only a few come from the Americas, Australia and Europe, as can be seen in Figure 10.1.

As soon as the location of geographic objects is involved ("where?"), a map becomes useful. However, maps can do more than just provide information on location. They can also inform about the thematic attributes of the geographic objects to be found in them. An example would be "What is the predominant land use in southeast Twente?" The answer could, again, be just verbal and state "Urban." However, such an answer does not reveal patterns of land use. In Figure 10.2, a dominant northwest–southeast urban buffer can be clearly distinguished. Maps can answer the "What?" question only in relation to location (the map as a reference frame). A third type of question that can be answered from maps is related to "When?" For instance, "When did the Netherlands have its longest coastline?" The answer might be "1600," and this would

probably be satisfactory for most people. However, it might be interesting to see how this has changed over the years. A set of maps as demonstrated in Figure 10.3 could provide the answer.

To summarize, maps can deal with questions/answers related to the basic components of spatial or geographic data: location (geometry), characteristics (thematic attributes) and time, and combinations thereof. As such, maps are the most efficient and effective means of transferring spatial information. The map user can locate geographic objects, while the shape and colour of signs and symbols representing the objects inform about their characteristics. They reveal spatial relations and patterns and offer the user insight into and overview of the distribution of particular phenomena. An additional characteristic of on-screen maps is that these are often interactive and have a link to a database; this allows for more complex queries.
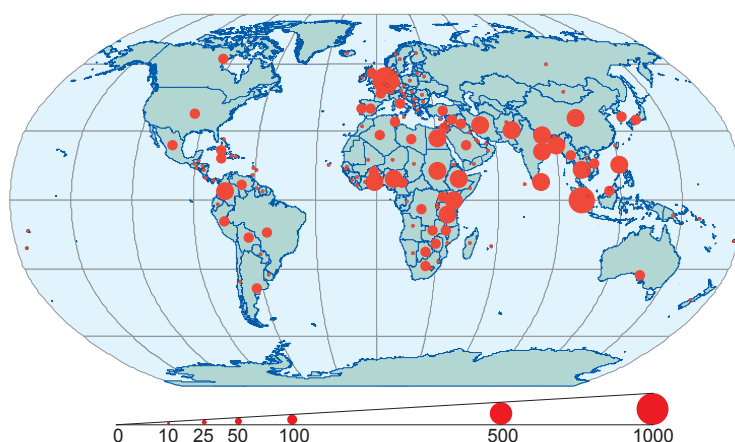
Figure 10.1
Maps and location: "Where do ITC cartography students come from?" Map scale is 1:200,000,000.

map as a model

Looking at these three sets of maps demonstrates an important quality of maps: the ability to offer an abstraction of reality. A map simplifies by leaving out certain details, but at the same time, when well designed, it puts the remaining information in a clear perspective. The map in Figure 10.1 only needs the boundaries of countries and a symbol to represent the number of students per country. In this particular case, there is no need to show cities, mountains, rivers or other phenomena.

This characteristic is well illustrated when one puts a map next to an aerial photograph or satellite image of the same area. Products such as the latter give all information observed by the capturing devices used. Figure 10.4 shows an aerial photograph of the ITC building and a map of the same area. The photograph shows all visible objects, including parked cars and small temporary buildings. From the photograph, it becomes clear that the weather, as well as the time of day, has had an influence on its contents: the shadow to the north of the buildings obscures other information. The map on the other hand, only gives the outlines of buildings and the surrounding streets. It is easier to interpret because of the selection/omission and classification of features. The symbolization chosen highlights the ITC building. Additional information, not available in the photograph, has been added, such as the name of the main street: Hengelosestraat. Other non-visible data, such as cadastral boundaries or even the sewerage system, could have been added in the same way. However, this also demonstrates that selection means interpretation, and there are subjective aspects to that. In certain circumstances, a combination of photographs and map elements can be useful.

There is a relationship between the effectiveness of a map for a given purpose and the
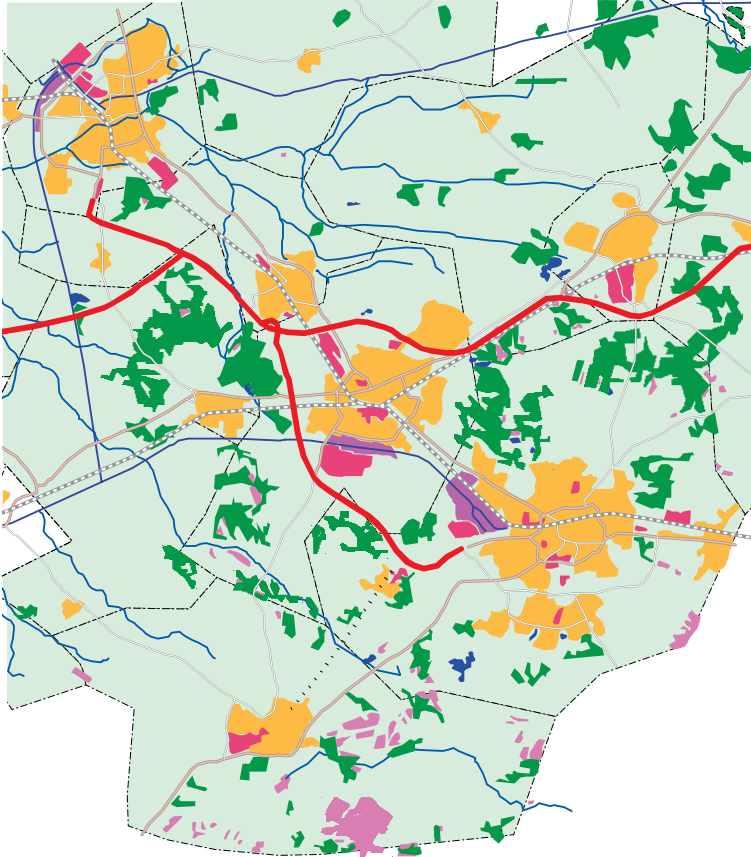
Figure 10.2
Maps and characteristics:
"What is the predominant
land use in southeast
Twente?"
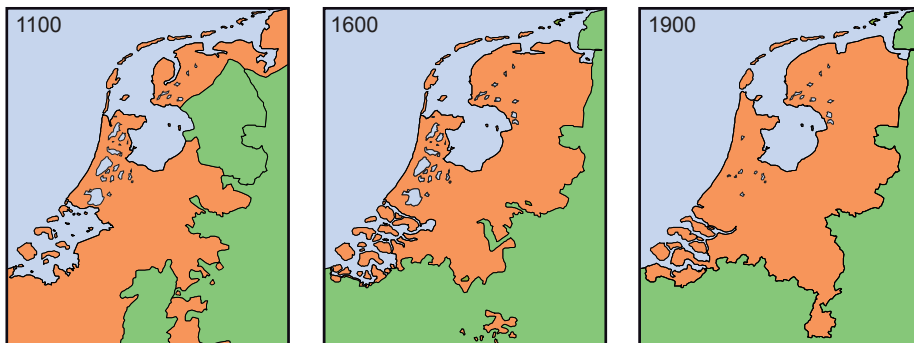
Figure 10.3
Maps and time: "When did
the Netherlands have its
longest coastline?"

map's scale. The Public Works department of a city council cannot use a 1:250,000 map for replacing broken sewer pipes, and the map of Figure 10.1 cannot be reproduced at scale 1:10,000. The map scale is the ratio between a distance on the map and the corresponding distance in reality. Maps that show much detail of a small area are called large-scale maps. The map in Figure 10.4 displaying the surroundings of the ITC building is an example of such a map. The world map in Figure 10.1 is a small-scale map. Scale indications on maps can be given verbally, such as "one-inch-to-the-mile", or as a representative fraction like 1:200,000,000 (1 cm on the map equals 200,000,000 cm (or 2000 km) in reality), or by a graphic representation such as the

scale bar on the map in Figure 10.4b. The advantage of using scale bars in digital environments is that its length also changes when the map is zoomed in, or enlarged, before printing. Sometimes it is necessary to convert maps from one scale to another, which may lead to problems of cartographic generalization.

<div style="margin-left: -1em; color: teal;">map definition</div>

Having discussed several characteristics of maps, we are now able to define a map. Board [10] defines a map as "a representation or abstraction of geographic reality. A tool for presenting geographic information in a way that is visual, digital or tactile." The first sentence in this definition contains three key words. The "geographic reality" represents the object of study: i.e. our world. "Representation" and "abstraction" refer to models of these geographic phenomena. The second sentence reflects the appearance of the map. Can we see or touch it? Or is it stored in a database? In other words, a map is a reduced and simplified representation of the Earth's surface, or parts of it, on a plane.

Figure 10.4
Comparing an aerial photograph (a) and a map (b) of the same area.

Traditionally, maps have been divided into two sorts: topographic maps and thematic maps. A topographic map visualizes, limited by its scale, the Earth's surface as accurately as possible. This may include infrastructure (e.g. railways and roads), land use (e.g. vegetation and built-up areas), relief, hydrology, geographic names and a reference grid. Figure 10.5 shows a small-scale topographic map (with text omitted) of Overijssel, the Dutch province in which Enschede is located.

Thematic maps represent the distribution of particular themes. One can distinguish between socio-economic themes and physical themes. The map in Figure 10.6a, showing population density in Overijssel, is an example of the first and the map in Figure 10.6b, displaying the province's drainage areas, is an example of the second. As can be observed, both thematic maps also contain some information found in the topographic map (Figure 10.5), so as to provide a geographic reference to the theme represented.

The amount of topographic information required depends on the map theme. In general, a physical map will need more topographic data than most socio-economic maps, which normally only need administrative boundaries. The map of drainage areas should have had rivers and canals added, while the inclusion of relief would also have made sense. Today's digital environment has diminished the distinction between topographic and thematic maps. Often, both topographic and thematic maps are stored in a database as separate data layers. Each layer contains data on a particular topic and the user is able to switch layers on or off at will.

The design of topographic maps is mostly based on conventions, some of which date back several centuries. Take, for example the following colour conventions: water in blue, forests in green, major roads in red, and urban areas in black. The design of
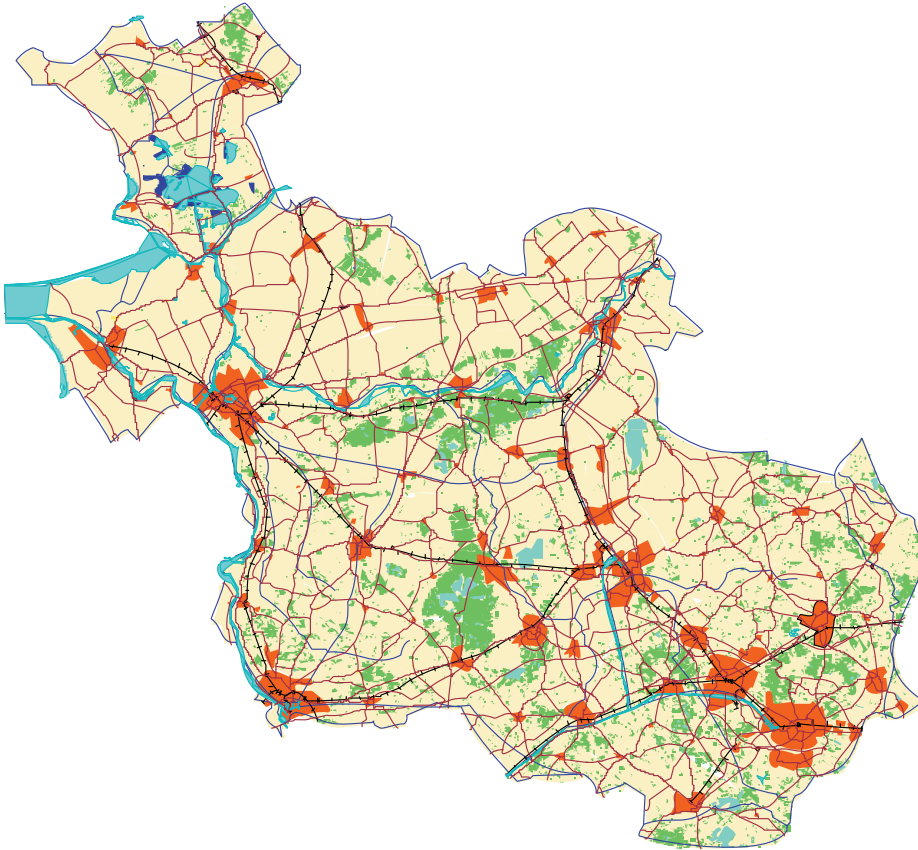
Figure 10.5
A topographic map of the province of Overijssel. Geographic names and a reference grid have been omitted for reasons of clarity.

Inhabitants per sq km

54-156
157-338
339-525
526-1972
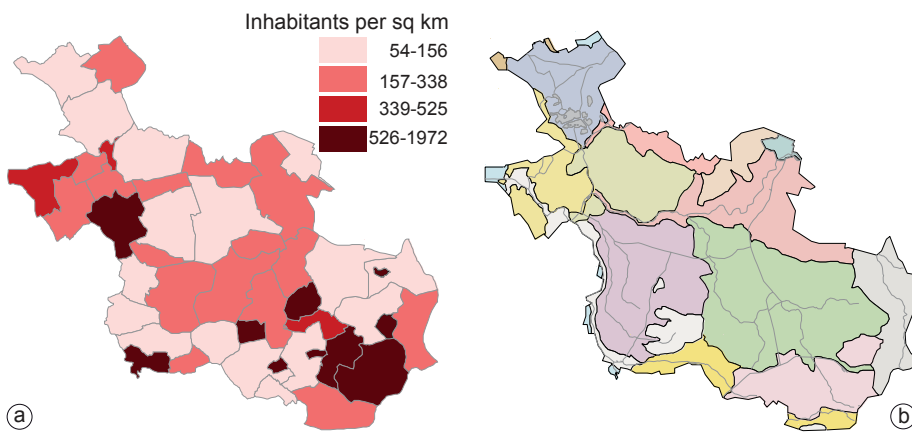
Figure 10.6
Thematic maps: (a) socio-economic thematic map, showing population density of the province of Overijssel (higher densities in darker tints); (b) physical thematic map, showing watershed areas of Overijssel.

thematic maps, however, should be based on a set of cartographic rules, also called *cartographic grammar*, which will be explained in Subsections 10.1.4 and 10.1.5 (but see also [60]).

cartographic grammar

Suppose that one wants to quantify land use changes in a certain area between 1990 and the current year. Two data sets (from 1990 and, say, 2008) can be combined with an overlay operation (see Section 9.4). The result of such a spatial analysis could be a spatial data layer from which a map can be produced to show the differences. The pa-

rameters used during the operation are based on models developed by the application at hand. It is easy to imagine that maps can play a role during this process of working with a GIS, by showing intermediate and final results of the GIS operations. Clearly, maps are no longer the only final product, which they used to be.

2D

3D

4D
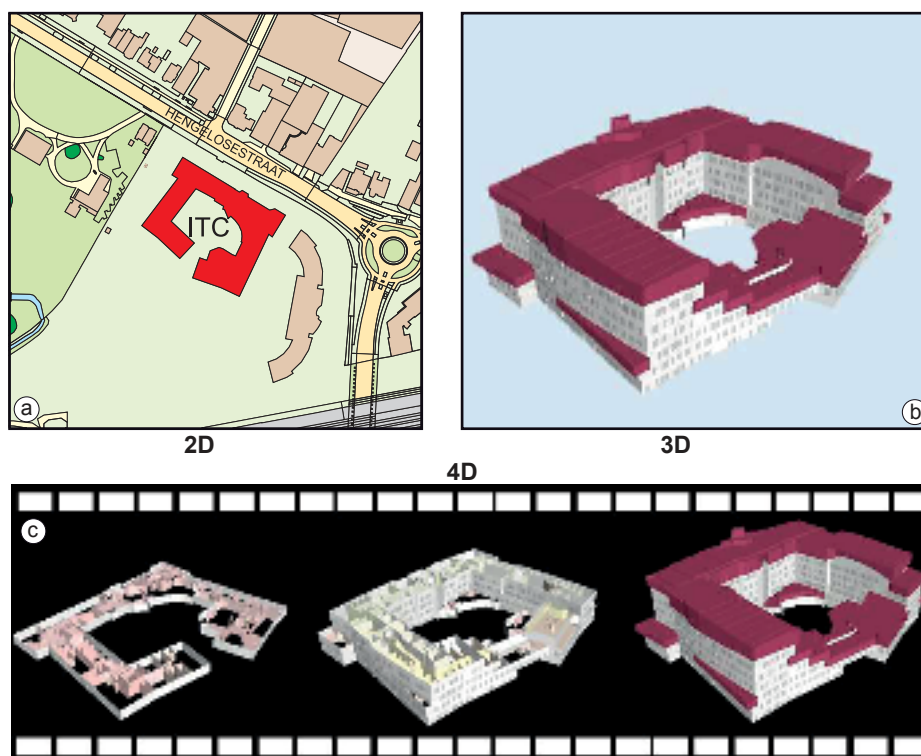
Figure 10.7
The dimensions of spatial data: (a) 2D, (b) 3D, (c) 3D with time.

Maps can further be distinguished according to the dimensions of spatial data that are graphically represented (Figure 10.7). GIS users also try to solve problems that deal with three-dimensional reality or with change processes. This results in a demand for other than just two-dimensional maps to represent geographic reality. Three-dimensional and even four-dimensional (namely, including time) maps are then required. New visualization techniques for these demands have been developed. Figure 10.7 shows the dimensionality of geographic objects and their graphic representation. Part (a) shows a map of the ITC building and its surroundings, while part (b) provides a three-dimensional view of the building. Figure 10.7c shows the effect of change, at three moments in time during the construction of the building.

10.1.2   The visualization process

The characteristic of maps and their function in relation to the spatial data handling process has been explained in the previous section. In this context, the cartographic visualization process is considered to be the translation or conversion of spatial data from a database into graphics, which are predominantly map-like products. During the visualization process, cartographic methods and techniques are applied. These can be considered to form a kind of grammar that allows for the optimal design and production of the maps, depending on the application (see Figure 10.8).

The producer of these visual products may be a professional cartographer, but they may also be an expert in a particular discipline, for instance someone mapping veg-

etation stands using remote sensing images or mapping health statistics in the slums of a city. To enable the translation from spatial data into graphics, we assume that the data are available and that the spatial database is well structured.
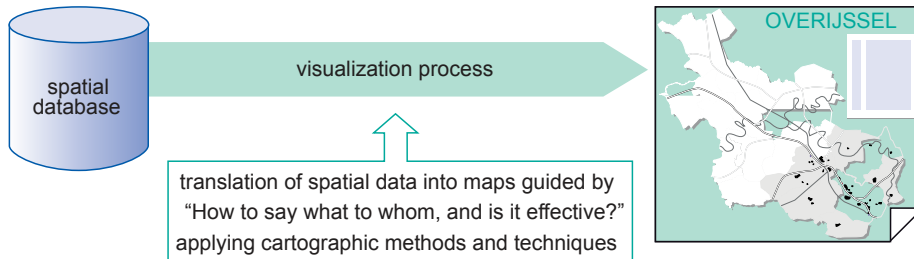
The visualization process can vary greatly depending on where in the spatial-data handling process it takes place and the purpose for which it is needed. Visualizations can be, and are, created during any phase of the spatial-data handling process as indicated before. They can be simple or complex, while the production time can be short or long.

Some examples are the creation of a full, traditional topographic map sheet, a newspaper map, a sketch map, a map from an electronic atlas, an animation showing the growth of a city, a three-dimensional view of a building or a mountain, or even a real-time map display of traffic conditions. Other examples include "quick and dirty" views of part of the database, the map used during the updating process or during a spatial analysis. However, visualization can also be used for checking the consistency of the acquisition process or even the database structure. These visualization examples from different phases in the process of spatial data handling demonstrate the need for an integrated approach to geoinformatics. The environment in which the visualization process is executed can vary considerably. It can be done on a stand-alone, personal computer, a network computer linked to an intranet, or on the World Wide Web (WWW/Internet).

In any of the examples just given, as well as for the maps in this book, the visualization process is guided by the question "How do I say what to whom?" "How" refers to cartographic methods and techniques; "I" represents the cartographer or map-maker; "say" deals with communicating in graphics the semantics of the spatial data; "What" refers to the spatial data and its characteristics, (for instance, whether they are of a qualitative or quantitative nature); "Whom" refers to the map audience and the purpose of the map—a map for scientists requires a different approach than a map on the same topic aimed at children. All these issues will be elaborated upon in following subsections.

In the past, cartographers were often solely responsible for the whole map compilation process. During this process, incomplete and uncertain data often still resulted in an authoritative map. The maps created by a cartographer had to be accepted by the user: cartography, for a long time, was very much driven by supply rather than demand. In some respects, this is still the case. However, nowadays one accepts that just making maps is not the only purpose of cartography. The visualization process should also be tested for its effectiveness. To the proposition "How do I say what to whom" we have to add "and is it effective?" Based on feedback from map users, or knowledge about the effectiveness of cartographic solutions, we can decide whether improvements are needed, and derive recommendations for future application of those solutions. In particular, with all the visualization options available today, for example animated maps, multimedia and virtual reality, it is essential to test the effectiveness of cartographic

methods and tools.

The visualization process is always influenced by several factors, i.e. the answers to questions:

- What will be the scale of the map: large, small, other? This introduces the problem of generalization. Generalization addresses the meaningful reduction of the map content during scale reduction.

- Are we dealing with topographic or thematic data? These two categories have traditionally resulted in different design approaches, as was explained in the previous subsection.

- More important for the design is the question of whether the data to be represented are of a quantitative or qualitative nature.

Some of these questions can be answered by just looking at the content of the spatial database.

We should understand that the impact of these factors/answers may increase, since the compilation of maps by spatial data handling is often the result of combining different data sets of different quality and from different data sources, collected at different scales and stored in different map projections. Cartographers have all kind of tools available to visualize the data. These tools consist of functions, rules and habits. Algorithms used to classify the data or to smooth a polyline are examples of functions. Rules tell us, for instance, to use proportional symbols to display absolute quantities or to position an artificial light source in the northwest to create a shaded relief map. Habits or conventions—or traditions as some would call them—tell us to colour the sea blue, lowlands green and mountains brown. The efficiency of these tools depends partly on the above-mentioned factors and partly on what map users are used to.

10.1.3   Visualization strategies: present or explore?

Traditionally, the cartographer's main task was the creation of good cartographic products. This is still true today. The main function of maps is to communicate geographic information, i.e. to inform the map user about location and the nature of geographic phenomena and spatial patterns. This has been the map's function throughout history. Well-trained cartographers are designing and producing maps, supported by a whole set of cartographic tools and theory as described in cartographic textbooks [105], [60].

Over the past few decades, many others have become involved in making maps. The widespread use of GISs has increased the number of maps made tremendously [69]. Even spreadsheet software commonly used in offices today has mapping capabilities, although most users are not aware of this. Many of these maps are not produced as final products, but rather as intermediaries to support the user in her/his work with spatial data. Hence, the map has started to play a completely new role: it is not only a communication tool, but also has become an aid in the user's (visual) thinking processes.

This thinking process is accelerated by continuing developments in hardware and software. New media such as CD-ROMs and the World Wide Web enable dynamic presentation and also user interaction. These have been accompanied by changing scientific and societal needs for georeferenced data and, as such, for maps. Users now expect immediate and real-time access to the data, data that have become abundant in many sectors of the geoinformation world. This abundance of data, seen as a "paradise" by some sectors, is a major problem in others. We lack the tools for user-friendly queries and retrieval when analysing the massive amount of (spatial) data produced

by sensors, which is now available via the Word Wide Web. A new branch of science is currently evolving to deal with this problem of abundance. In the geo-disciplines, it is called visual data mining.

All these developments have enhanced the meaning of the term visualization. According to dictionaries, it means "to make visible" or "to represent in graphical form". It can be argued that, in the case of spatial data, this has always been the business of cartographers. However, progress in other disciplines has linked the word to more specific ways in which modern computer technology can facilitate the process of "making visible" in real time. Specific software toolboxes have been developed, and their functionality is based on two key words: *interaction* and *dynamics*. A separate discipline called scientific visualization has developed around those keywords [72] and it has also had an important impact on cartography. New discipline offers the user the possibility of instantaneously changing the appearance of a map. Interaction with the map will stimulate the user's thinking and will add a new functions to the map: it not only communicates, but also prompts thinking and decision-making.

Developments in scientific visualization stimulated DiBiase [29] to define a model for map-based scientific visualization, also known as geovisualization, that covers both the presentation and exploration functions of the map (see Figure 10.9). Presentation is described as "public visual communication" since it concerns maps aimed at wide audiences. Exploration is defined as "private visual thinking" because it is often done by an individual playing with the spatial data to determine its significance. It is obvious that presentation fits into the traditional realm of cartography, where the cartographer works on known spatial data and creates communicative maps. Such maps are often created for multiple uses.

Exploration, however, often involves an expert in a particular discipline who creates maps while dealing with unknown data. These maps are generally for a single purpose, expedient in the expert's attempt to solve a problem. While dealing with the data, the expert should be able to rely on cartographic expertise provided by the software or some other means. Essentially, here the problem of translation of spatial data into cartographic symbols also needs to be solved. The above trends all have to do with what has been called by Morrison [77] the "democratization of cartography". As he explains it: "using electronic technology, no longer does the map user depend on what the cartographer decides to put on a map. Today the user is the cartographer . . . users are now able to produce analyses and visualizations at will to any accuracy standard that satisfies them."

Exploration means to search for spatial, temporal or spatio-temporal patterns, and relationships between patterns or trends. In the case of a search for patterns, a domain expert may be interested in aspects such as the distribution of a phenomenon, the occurrence of anomalies, or the sequence of appearances and disappearances. A search for relationships between patterns could include, for example, changes in vegetation indices and climatic parameters or locations of deprived urban areas and their distance to educational facilities. A search for trends could, for example, focus on developments in the distribution and frequency of landslides. Maps not only enable these types of searches; findings may also trigger new questions and lead to new visual exploration (or analysis).

What is unknown for one is not necessarily unknown to others. For instance, browsing in an atlas of the World on a DVD is an exploration for most of us, because of the wealth of information at our finger tips. With products like these, such exploration takes place within boundaries set by the producers. Browsing Google Earth is probably an adventure for everyone since everyone can add their own data! Cartographic knowledge is incorporated in the program, resulting in pre-designed maps. Some
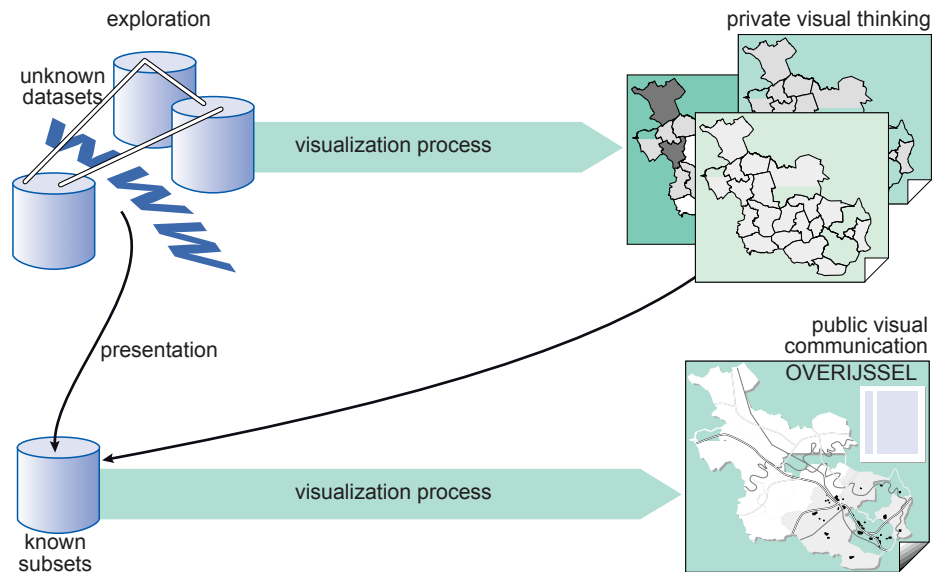
Figure 10.9
Private visual thinking and
public visual communication.

users may feel this to be a constraint, but the same users will probably no longer feel constrained as soon as they follow the web links attached to this electronic atlas. This shows that the data, the users, and the user environment influence one's view of what exploration entails.

To create a map, one selects relevant geographic data and converts these into meaningful symbols for the map. In the past, printed maps had a dual function. They acted as a database of the objects selected from reality, and they communicated information about those geographic objects. The introduction of computer technology and databases, in particular, has created a split between these two functions of the map. The database function is no longer required for the map, although each map can still function like a database. The communication function of maps has not changed.

The sentence "How do I say what to whom, and is it effective?" guides the cartographic visualization process and summarizes the cartographic communication principle. Especially when dealing with maps in the realm of presentation cartography (Figure 10.9), it is important to adhere to cartographic design rules. This is to guarantee that the resulting maps are easily understood by their users.

How does this communication process work? Well, see Figure 10.10. The process starts with information to be mapped (the "What" from the sentence). Before anything can be done, the cartographer should get a feel for the nature of the information, since this determines the graphical options; cartographic information analysis provides that feel. From this knowledge, the cartographer can choose the correct symbols to represent the information in the map. Cartographers have a whole toolbox of visual variables available to match symbols with the nature of the data. For the rules, see Subsection 10.1.4.

In 1967, the French cartographer Bertin published the basic concepts of the theory of map design in his book Sï£¡miologie Graphique [7]. He provided guidelines for making good maps. Nevertheless, if 10 professional cartographers were given the same mapping task and each were to apply Bertin's rules (see Subsection 10.1.4, this would still result in 10 different maps. For instance, if the guidelines dictate the use of colour, it is not stated which colours should be used. Still, all 10 maps could be of

good quality.

Returning to the scheme, the map (the medium that does the "say" in our sentence above) is read by the map users (the "whom" from the sentence). They extract some information from the map, represented by the box entitled "Info retrieved". From the figure it becomes clear that the boxes with "Information" and "Info retrieved" do not overlap. This means the information derived by the map user is not the same as the information that the cartographic communication process started with.
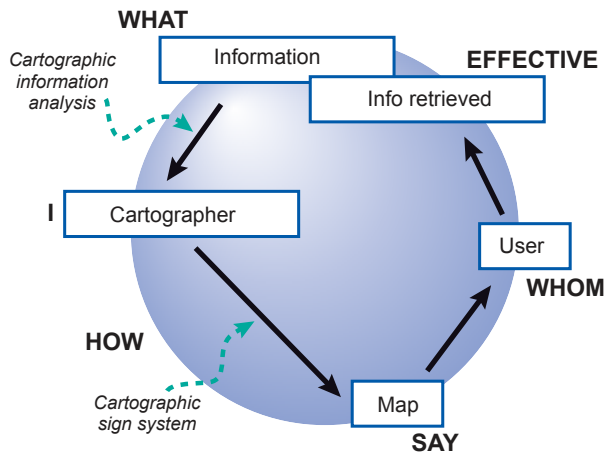
Figure 10.10
The cartographic communication process, based on "How do I say what to whom, and is it effective?"

There may be several causes for this. Perhaps all the original information was not used or perhaps additional information was added during the process. Perhaps the cartographer deliberately omitted information with the aim of emphasizing the remaining information. Another possibility is that the map user does not understand the map fully. Information gained during the communication process could be the result of the cartographer adding extra information to strengthen the information already available. It is also possible that the map user has some prior knowledge on the topic or area, which would allow him or her to combine this prior knowledge with knowledge retrieved from the map.

10.1.4 The cartographic toolbox

What kind of data do I have?

To find the proper symbology for a map one has to analyse the cartographic data. The core of this process of analysis is to access the characteristics of the data to find out how they can be visualized, so that the map user will interpret them properly. The first step in the analysis process is to find a common denominator for all the data. This common denominator will then be used as the title of the map. For instance, if all data are related to land use collected in 2005, the title could be "Land use of . . . 2005". Second, the individual component(s), such as land use, and probably relief, should be analysed and their nature described. Later, these components should be visible in the map legend.

We have already discussed different kinds of data values in Section 8.1 in relation to the types of computations we can do on them. Now it is time to look at the different types of data in relation to how they might be mapped or displayed.

Data will be of a qualitative or quantitative nature. Qualitative data is also called nominal data, which exists as discrete, named values without a natural order amongst the values. Examples are different languages (e.g. English, Swahili, Dutch), different soil

data types

types (e.g. sand, clay, peat) or different land use categories (e.g. arable land, pasture). In the map, qualitative data are classified according to disciplinary insights, such as a soil classification system represented as basic geographic units: homogeneous areas associated with a single soil type, recognizable by the soil classification.

Quantitative data can be measured, either along an interval or ratio scale. For data measured on an interval scale, the exact distance between values is known, but there is no absolute zero on the scale. Temperature is an example: $40\,°C$ is not twice as hot as $20\,°C$, and $0\,°C$ is not an absolute zero.

Quantitative data with a ratio scale do have a known absolute zero. An example is income: someone earning $100 earns twice as much as someone with an income of $50. In order to generate maps, quantitative data are often classified into categories according to some mathematical method.

In between qualitative and quantitative data, one can distinguish ordinal data. These data are measured along a relative scale and are as such based on hierarchy. For instance, one knows that a particular value is "more" than another value, such as "warm" versus "cool". Another example is a hierarchy of road types: "highway", "main road", "secondary road" and "track". The different types of data are summarized in Table 10.1.

Table 10.1
Differences in the nature of data and their measurement scales.

Measurement scale	Nature of data
Nominal, categorical	Data of different nature / identity of things (qualitative)
Ordinal	Data with a clear element of order, though not quantitatively determined (ordered)
Interval	Quantitative information with arbitrary zero
Ratio	Quantitative data with absolute zero

How can I map my data?

Basic elements of a map, irrespective of the medium on which it is displayed, are point symbols, line symbols, area symbols, and text. The appearance of point, line, and area symbols can vary depending on their nature. Most maps in this book show symbols in different size, shape and colour. Points can vary in form or colour to represent the location of shops or they can vary in size to represent aggregated values (e.g. number of inhabitants) for an administrative area. Lines can vary in colour to distinguish between administrative boundaries and rivers, or vary in shape to show the difference between railroads and roads. Areas follow the same principles: differences in colour distinguishes between different vegetation.

Although variations in the appearance of symbols are only limited by the imagination, there are a few categories into which they can be grouped. Bertin [7] distinguished six categories, which he called the visual variables, which may be applied to point, line and area symbols. As illustrated in Figure 10.11, the symbols are:

map symbology

- size;

- value (lightness);

- texture;

- colour;

- orientation; and

- shape.

These visual variables can be used to make one symbol different from another. In doing this, map-makers have, in principle, freedom of choice, provided they do not violate the rules of cartographic grammar. They do not have any such choice when deciding where to locate the symbol in the map: the symbol should be located where the feature belongs. Visual variables influence the map user's perception in different ways. What is perceived depends on the human capacity to see:

- what is of equal importance (e.g. all red symbols represent danger), saturation differences;

- order (e.g. the population density varies from low to high—represented by light and dark colour tints, respectively);

- quantities (e.g. symbols changing in size with small symbols for small amounts); and

- instant overview of the mapped theme.
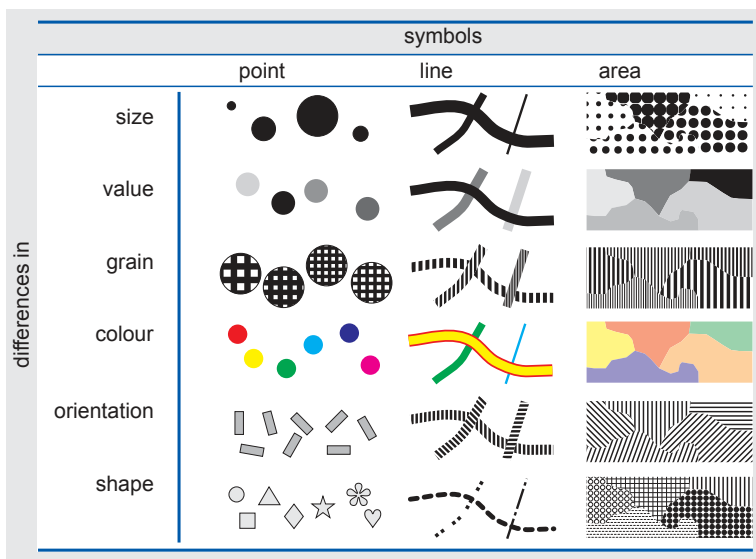
Figure 10.11
Bertin's six visual variables illustrated.

There is an obvious relationship between the nature of the data to be mapped and the "perception properties" of visual variables. In Table 10.2, the measurement scales as defined in Table 10.1 are linked to the visual variables displayed in Figure 10.11. "Dimensions of the plane" is added to the list of visual variables; it is the basis, used for the proper location of symbols on the plane (map). The perception properties of the remaining visual variables have been added. In the next subsection we discuss some typical mapping problems and demonstrate the use of the principles that have been outlined.

Perception properties	Visual variables	Measurement scales			
		Nominal	Ordinal	Interval	Ratio
	Dimensions of the plane	x	x	x	x
Order & quantities	Size		x	x	x
Order	(Grey) value		x	x	
	Grain/texture		x	x	
Equal importance	Colour hue	x			
	Orientation	x			
	Shape	x			

10.1.5 How to map ...?

This subsection deals with characteristic mapping problems. We first describe a problem and then briefly discuss a solution based on cartographic rules and guidelines. The need to follow these rules and guidelines is illustrated by some maps that have been poorly designed—but are nevertheless commonly found.
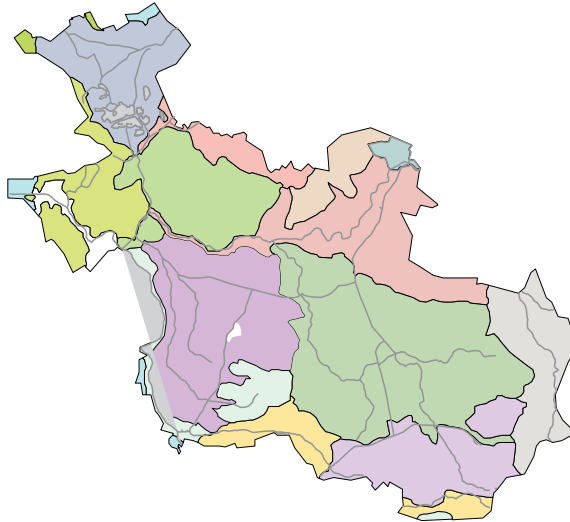
How to map qualitative data

If, after a long period of fieldwork, someone has finally delineated the boundaries of a province's watersheds, probably they will be interested in making a map showing these areas. The geographic units in the map will have to represent the individual watersheds. In such a map, each of the watersheds should get equal attention; none should stand out above the others.

The application of colour would be the best solution since is has characteristics that allow one to quickly differentiate between different geographic units. However, since none of the watersheds is more important than the others, the colours used have to be of equal visual weight or brightness. Figure 10.12 gives an example of a map in which colour has been used correctly. The readability is influenced by the number of displayed geographic units. In this example, there are about 15. When this number is much higher, the map, at the scale displayed here, will become too cluttered. The map can also be made by depicting the watershed areas with different forms (small circles, squares, triangles, etc.) in one colour (e.g. black for a monochrome map)—as an application of the visual variable "shape". The amount of geographic units that can be displayed is then even more critical.

Figure 10.13 shows two examples of how not to create such a map. In (a), several tints of black are used—as an application of the visual variable "value". Looking at the map may cause perceptual confusion since the map image suggests differences in importance that are not there in reality. In Figure 10.13b, colours are used instead. However, where most watersheds are represented in pastel tints, one of them stands out because of its bright colour. This gives the map an unbalanced look. The viewer's eye will be distracted by the bright colours, resulting in unjustified weaker attention for other areas.

How to map quantitative data

If, after executing a census, one would like to create a map of the number of people living in each municipality, one would be dealing with absolute quantitative data. The geographic units will logically be the municipalities. The final map should allow the user to determine the amount per municipality and also offer an overview of the geo-

graphic distribution of the phenomenon. To achieve this objective, the symbols used
should possess properties that facilitate quantitative perception. Symbols varying in
size would fulfill this demand. Figure 10.14 shows the final map of inhabitants in the
province of Overijssel.
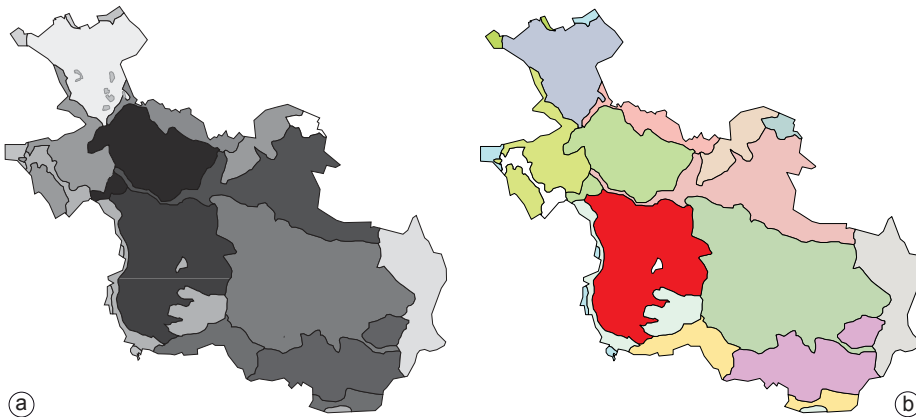
The fact that it is easy to make errors can be seen in Figure 10.15. In Figure 10.15a, dif-
ferent tints of green (the visual variable "value") have been used to represent absolute
population numbers. The reader might get a reasonable impression of the individual
amounts but not of the actual geographic distribution of the population, as the size of
the geographic units will influence the perceptional properties too much. Imagine a
small and a large unit having the same number of inhabitants. The large unit would
visually attract more attention, giving the impression there are more people than in
the small unit. Another issue is that the population is not necessarily homogeneously
distributed within the geographic units.

Colour has also been misused in Figure 10.15b. The four-colour scheme applied makes
it is impossible to infer whether red represents more-populated areas than blue. It
is impossible to instantaneously answer a question like "Where do most people in
Overijssel live?"
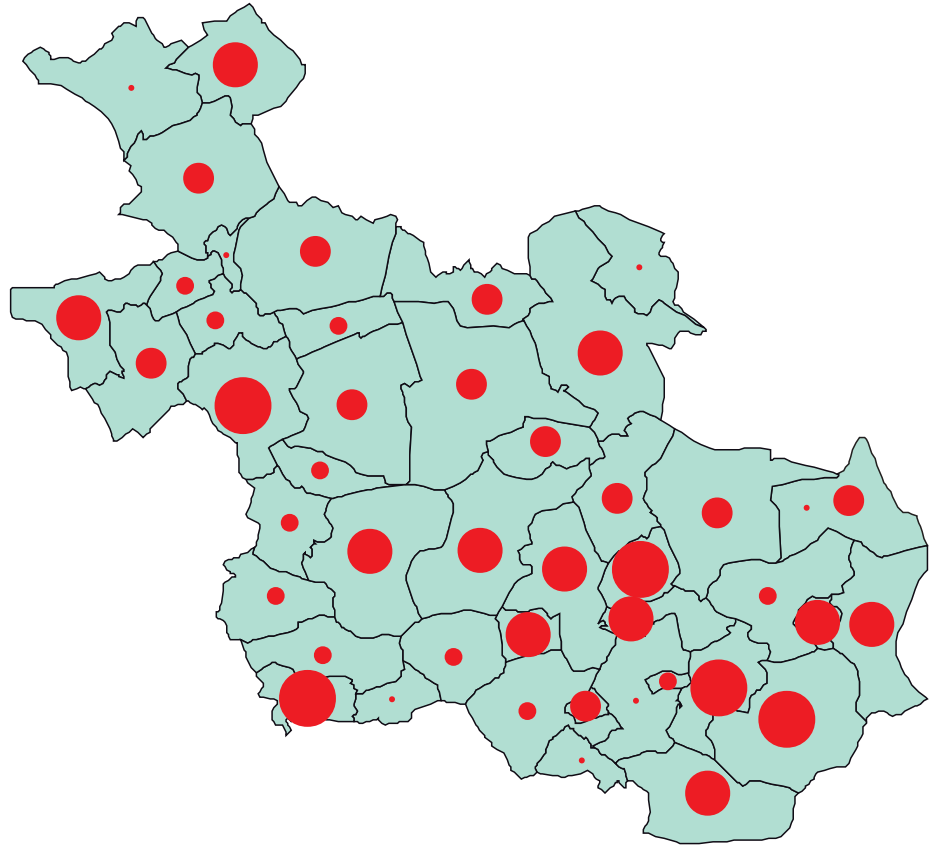
Figure 10.14
Mapping absolute
quantitative data.
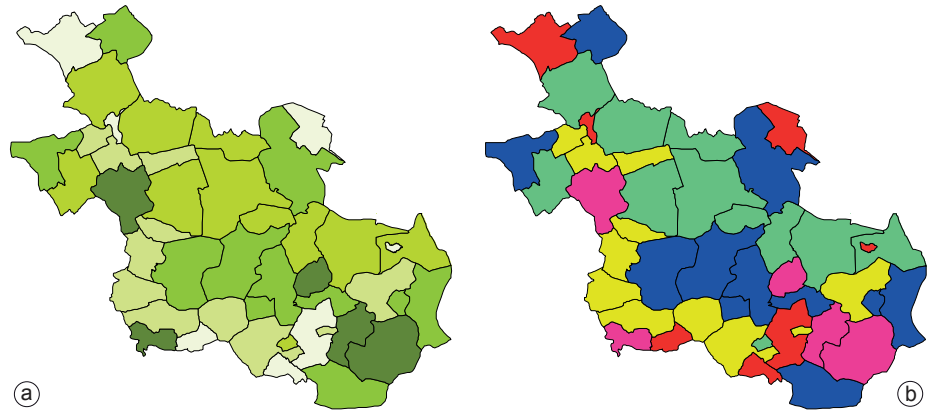
Figure 10.15
Poorly designed maps
displaying absolute
quantitative data: (a)
incorrect use of green tints for
absolute population figures;
(b) incorrect use of colour.

On the basis of absolute population numbers per municipality and their geographic size, we can also generate a map that shows population density per municipality. We are then dealing with relative quantitative data. The numbers now have a clear relation with the area they represent. The geographic unit will again be "municipality". The aim of the map is to give an overview of the distribution of the population density. In the map of Figure 10.16, value has been used to display the density from low (light tints) to high (dark tints). The map reader will automatically, at a glance, associate the

dark colours with high density and the light colours with low density.

Figure 10.17a shows the effect of incorrect application of the visual variable "value". In this map, the value tints are out of sequence. The user has to go to quite some trouble to find out where in the province the high-density areas can be found. Why should the lighter, mahogany-red represent areas with a higher population density than darker, burgundy-red. In Figure 10.17b colour has been used in combination with value. The first impression of the map reader would be to think that the brown areas represent those areas with the highest density. A closer look at a legend would tell you that this is not the case and that those areas are represented by another colour that does not "speak for itself".

If one studies the poorly designed maps carefully, the information can be derived in one way or another, but it takes quite some effort to do so. Proper application of cartographic guidelines guarantee that this will go much more smoothly (e.g. faster and with less chance of misunderstanding).

Inhabitants per sq km

54-156
157-338
339-525
526-1972
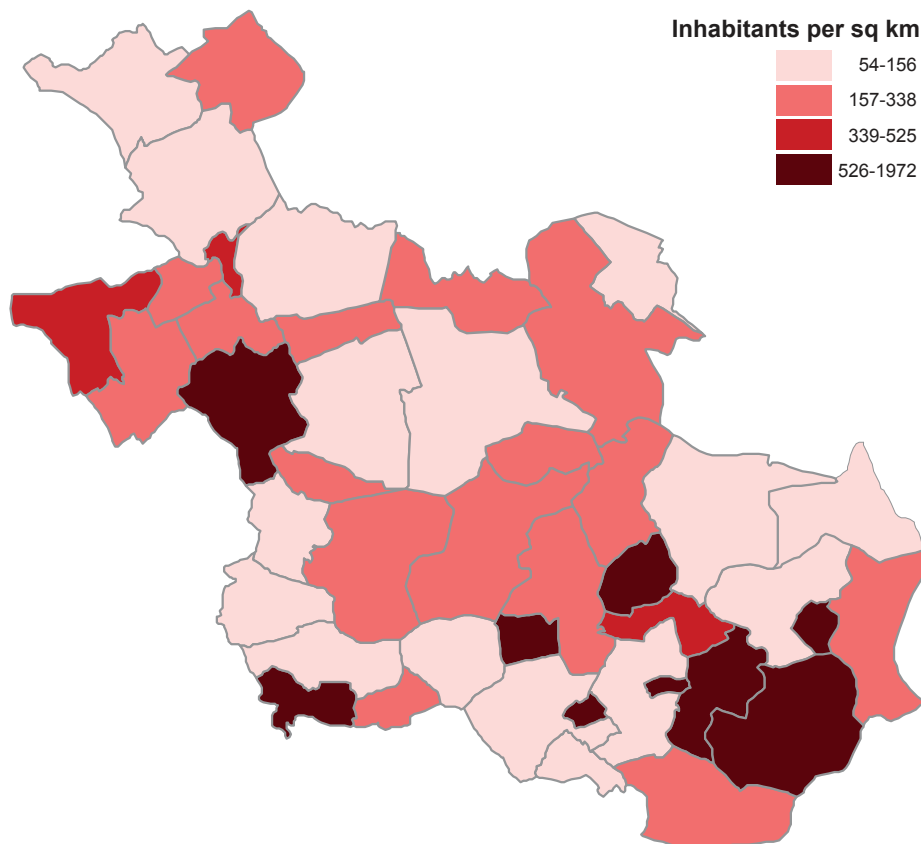
Figure 10.16
Mapping relative quantitative data.

How to map terrain elevation

Terrain elevation can be mapped using various methods. Often, one will have collected an elevation data set for individual points such as peaks, or other characteristic points in the terrain. Obviously, one can map the individual points and add the height information as text. However, a contour map, in which the contour lines connect points of equal elevation, is generally used. To visually improve the information content of such a map, the space between the contour lines can be filled with colour

Inhabitants per sq km

	54-156
	157-338
	339-525
	526-1972
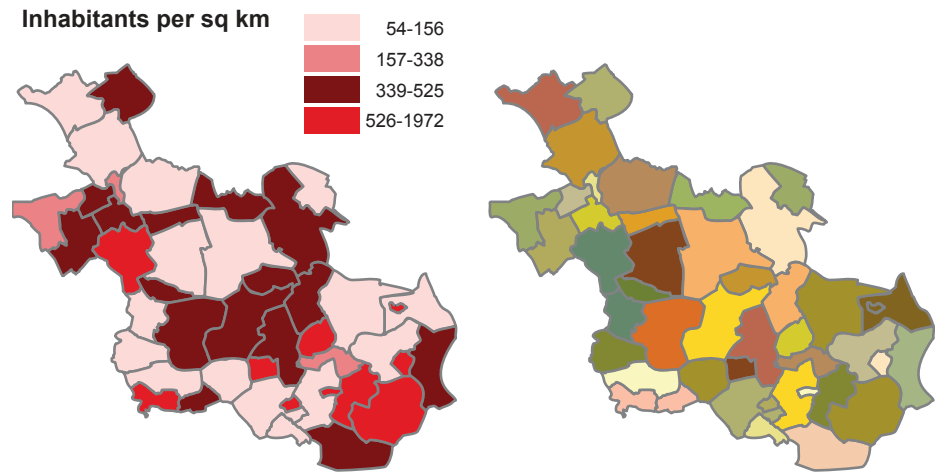
Figure 10.17
Poorly designed maps representing relative quantitative data: (a) lightness values used out of sequence; (b) colour should not be used.

and value information following a convention, e.g. green for low elevation and brown for areas of high elevation. This technique is known as hypsometric or layer tinting.

Even more advanced is the addition of shaded relief. This will improve the impression of three-dimensional relief (see Figure 10.18). The shaded-relief map uses all three-dimensional information available to create shading effects. This map, represented on a two-dimensional surface, can also be floated in three-dimensional space to give it the real three-dimensional appearance of a "virtual world", as shown in Figure 10.18d. Looking at such a representation, one can immediately imagine that it will not always be effective. Certain (low) objects in the map will easily disappear behind other (higher) objects. Interactive functions are required to manipulate the map in three-dimensional space in order to look behind some objects. These manipulations include panning, zooming, rotating and scaling.

Scaling is needed particularly along the z-axis since some maps require small-scale elevation resolution, while other require large-scale resolution, i.e. vertical exaggeration. One can even imagine that other geographic, three-dimensional objects (for instance, the built-up area of a city and individual houses) have been placed on top of the terrain model, as is done in Google Earth. Of course, one can also visualize objects below the surface in a similar way, but this is more difficult because the data to describe underground objects are sparse.

Socio-economic data can also be viewed in three dimensions. This may result in dramatic images that will be long remembered by the map user. Figure 10.19 shows the absolute population figures of Overijssel in three dimensions. Instead of using proportionally sized circles to depict the number of people living in a municipality (as we did in Figure 10.14), the proportional height of a municipality now indicates total population. The image clearly shows that the municipality of Enschede (the largest column to the lower right) has by far the highest population.

How to map time series

The third dimension of GIS routines is no longer reserved solely for advanced handling of spatial data. Nowadays, the handling of time-dependent data is also a feature of these routines, a consequence of the increasing availability of data captured at different periods in time. In addition to this abundance of data, the GIS community wants to analyse changes caused by real-world processes. To that end, single time-slices of data are no longer sufficient and the visualization of these processes cannot
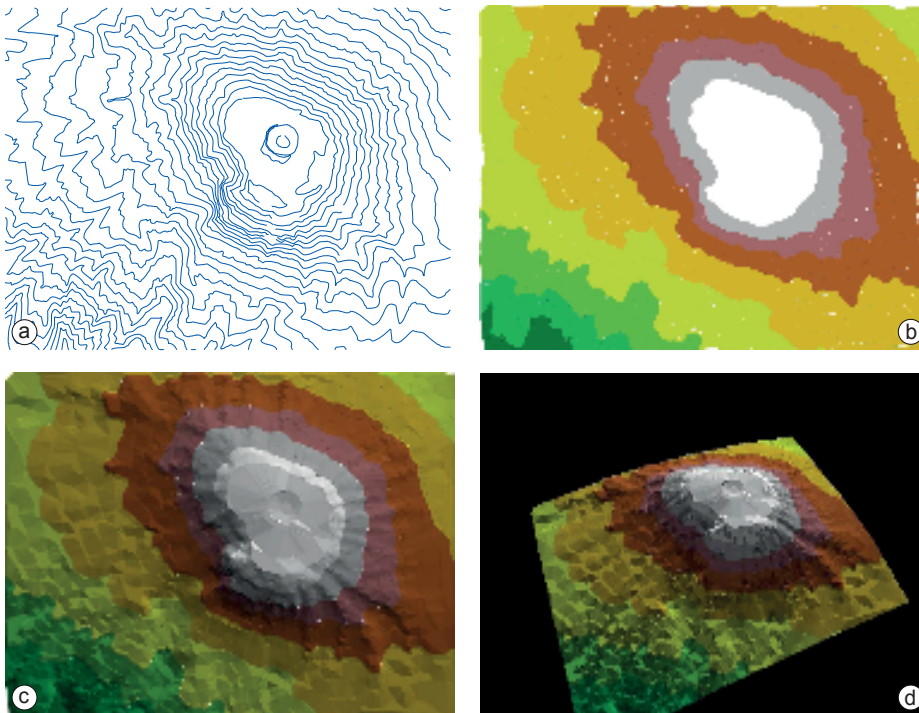
be supported by only static, printed maps.

Mapping time means mapping change. This may be change in a feature's geometry, in its attributes, or both. Examples of changing geometry are the evolving coastline of the Netherlands (as displayed in Figure 10.3), the location of Europe's national boundaries, or the position of weather fronts. Changes in the ownership of a land parcel, in land use or in road traffic intensity are other examples of changing attributes. Urban growth is a combination of both: urban boundaries expand with growth and simultaneously land use shifts from rural to urban. If maps are to represent events like these, they should be suggestive of such change.

Suggestion of change implies the use of symbols that are perceived as representing change. Examples of such symbols are arrows that have an origin and a destination. These are used to show movement and their size can be an indication of the magnitude

of change. Size changes can also be applied to other point and line symbols to show increase and decrease over time. Specific point symbols such as "crossed swords" (battle) or "lightning" (riots) can be found to represent dynamics in historic maps. Another alternative is the use of the visual variable value (expressed as tints). In a map showing the development of a town, dark tints represent old built-up areas, while new built-up areas are represented by light tints (see Figure 10.20a).
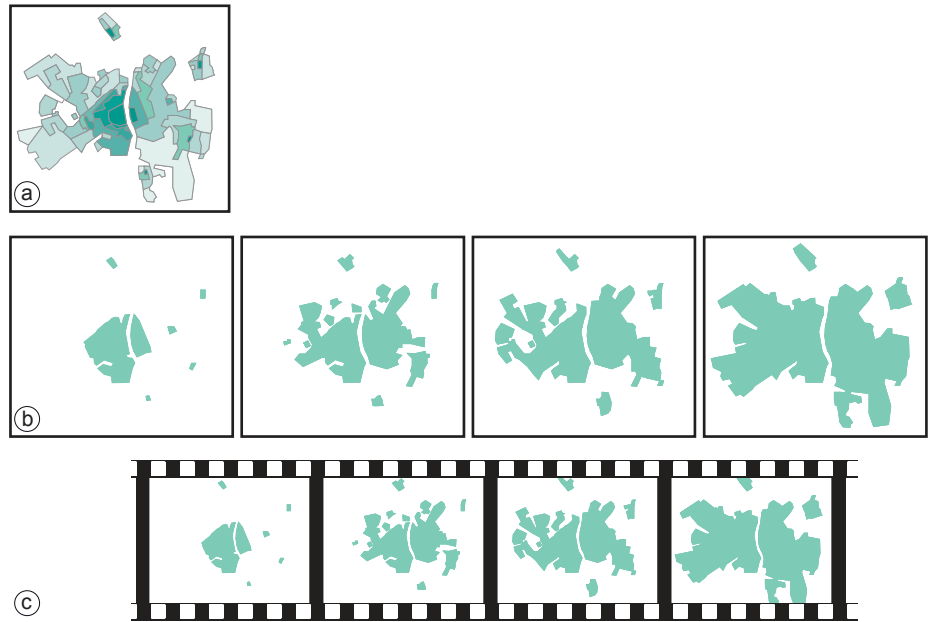
Figure 10.20
Mapping change; example of the urban growth of the city of Maastricht, The Netherlands: (a) single map, in which tints represent age of the built-up area; (b) series of maps; (c) (simulation of an) animation.

Three temporal cartographic techniques can be distinguished (see Figure 10.20):

- *Single static map* Specific graphic variables and symbols are used to indicate change or represent an event. Figure 10.20a applies the visual variable "value" to represent the age of built-up areas;

- *Series of static maps* A single map in the series represents a "snapshot" in time. Together, the maps depict a process of change. Change is perceived by the succession of individual maps depicting the situation in successive snapshots. It could be said that the temporal sequence is represented by a spatial sequence that the user has to follow to perceive the temporal variation. The number of images should be limited since it is difficult for the human eye to follow long series of maps (Figure 10.20b);

- *Animated map* Change is perceived to evolve in a single image by displaying several snapshots one after the other, just like a video clip of successive frames. The difference from the series of maps is that the variation can be deduced from real "change" seen taking place in the image itself, not from a spatial sequence (Figure 10.20c). For the user of a cartographic animation, it is important to have tools available that allow for interaction while viewing the animation. Seeing an animation play will often leave users with many questions about what they have seen. And just replaying the animation is not sufficient to answer questions like "What was the position of the northern coastline during the 15th century?" Most of the general software packages for viewing animations already offer facilities

such as "pause" (to look at a particular frame) and '(fast-)forward' and '(fast-)backward', or step-by-step display. More options have to be added, such as the possibility to go directly to a certain frame based on a task command like: "Go to 1850".

10.1.6 Map cosmetics

Most maps in this chapter are correct from a cartographic grammar perspective. However, many of them lack the additional information needed to be fully understood that is usually placed in the margin of printed maps. Each map should have, next to the map image, a title to inform the user of the topic visualized. A legend is necessary to understand how the topic is depicted. Additional marginal information to be found on a map is a scale indicator, a North arrow for orientation, the map datum and map projection used, and some lineage information, (such as data sources, dates of data collection and methods used). Furthermore, information can be added that indicates when the map was issued and by whom (author/publisher).

All this information allows the user to obtain an impression of the quality of the map and is comparable with meta-data describing the contents of a database or data layer. Figure 10.21 illustrates these map elements. On printed maps, these elements (if all are relevant) have to appear next to the map face itself. Maps presented on screen often do without marginal information, partly because of space constraints. However, on-screen maps are often interactive and clicking on a map element may reveal additional information from the database. Legends and titles are often also available on demand.
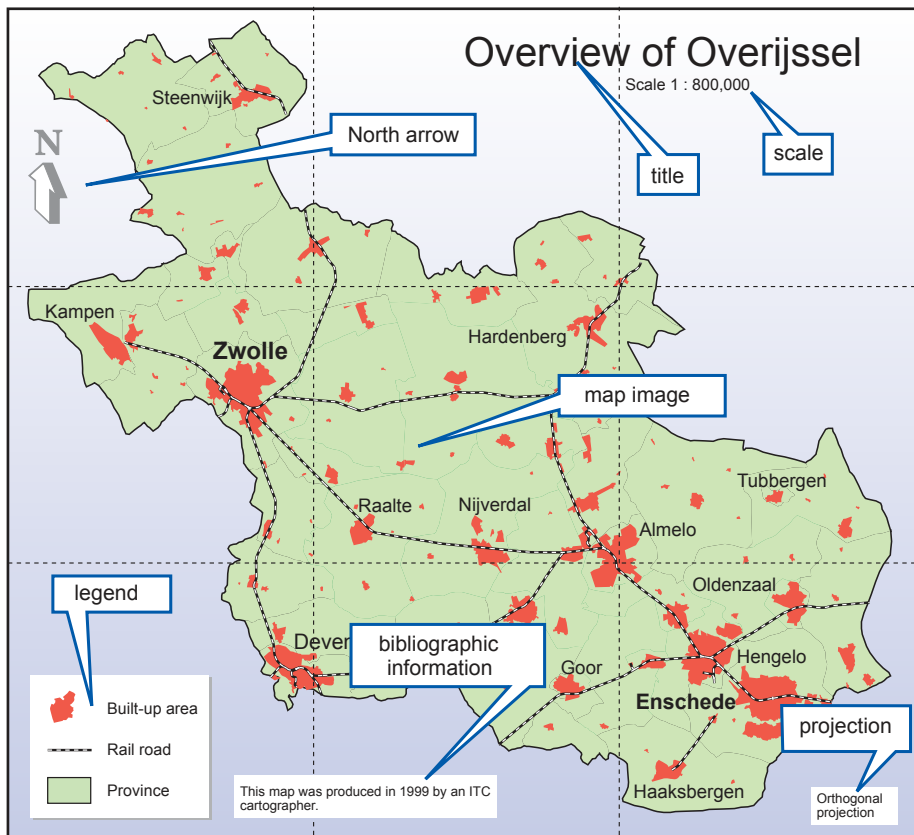
Figure 10.21
A printed map and its marginal information.

The map in Figure 10.21 is one of the first in this section that has text included. Figure 10.22 is another example. Text is used to transfer information in addition to that through the symbols used. This can be done by the applying visual elements to the text as well, as in Figure 10.22: italics—is a case of the visual variable orientation—have been used for building names to distinguish them from road names. Another common example is the use of colour to differentiate (at a nominal level) between hydrographic names (in blue) and other names (in black). The text should also be placed in a proper position with respect to the object to which it refers.

Figure 10.22
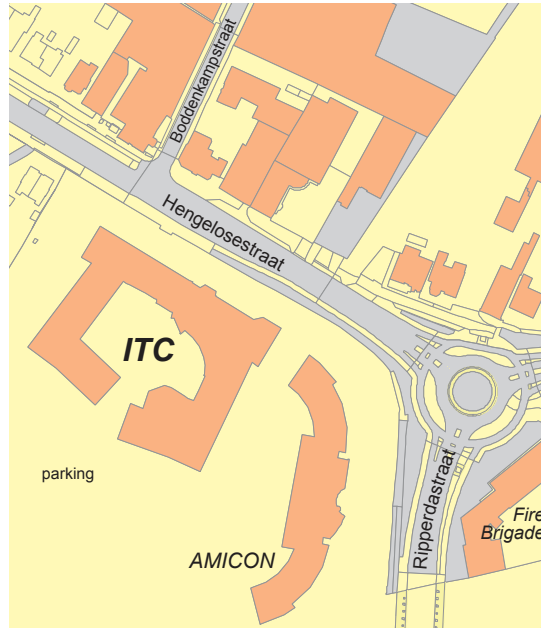Text in a map.

Figure 10.23
Visual hierarchy and the location of the ITC building: (a) hierarchy not applied; (b) hierarchy applied.

Maps constructed according to basic cartographic guidelines are not necessarily visually appealing maps. Although well-constructed, they might still look sterile. The design aspect of creating appealing maps also has to be included in the visualization

process. "Appealing" does not only mean having nice colours. One of the keywords here is contrast. Contrast will increase the communication role of the map since it creates a hierarchy in the map contents, assuming that not all information is of equal importance. This design trick is known as visual hierarchy or the figure–ground concept. The need for visual hierarchy in a map is best understood by looking at the map in Figure 10.23a, which just shows lines. The map of the ITC building and surroundings in part (b) is an example of a map that has visual hierarchy applied. The first object to be noted will be the ITC building (the darkest patch in the map) followed by other buildings, with the road on a lower level and the urban land parcels at the lowest level.

10.1.7 Map dissemination

Map design will not only be influenced by the nature of the data to be mapped or the intended audience (the "what" and "whom" from "How do I say what to whom, and is it effective?"). The output medium also plays a role. Traditionally, maps were produced on paper, and many still are. Currently, though, many maps are presented on-screen, for a quick look, for a presentation or for display on the Internet. Compared to maps on paper, on-screen maps have to be smaller, so their contents should be carefully selected. This might seem a disadvantage, but presenting maps on-screen offers very interesting alternatives. In the previous subsection, we mentioned that the legend only needs to be a mouse click away. A mouse click could also open a link to a database, revealing much more information than a printed map could ever offer. Links to other than tabular or map data could also be made available.

Maps and multimedia (photography, sound, video, animation) can be integrated. Some of today's web atlases are good examples of how multimedia elements can be integrated with a map. For example, pointing to a country on a world map may start the national anthem of the country or shows its flag. It can be used to explore a country's language; moving the mouse would start a short sentence in the region's dialects.

The World Wide Web is nowadays a medium commonly used to present and disseminate spatial data. Maps can, however, still play their traditional role, for instance to show the location of objects or provide insight into spatial patterns, but because of the nature of the Internet, a map can also function as an interface to additional information. Geographic locations on the map can be linked to photographs, text, sound or other maps, perhaps even functions such as on-line booking services.

Maps can also be used as previews of spatial-data products to be acquired through a spatial-data clearinghouse that is part of a Spatial Data Infrastructure. For that purpose, we can make use of geo-webservices, which can provide interactive map views as an intermediate step between data and web browser (please refer to Subsection 8.5.4).

How can maps be used on the Internet? We can distinguish several methods that differ in terms of technical skills needed from both the user's and provider's perspective. An important distinction is the one between static and dynamic maps.

Many static maps on the Web are view-only. Organizations, such as map libraries or tourist information providers, often make their maps available in this way. This form of presentation can be very useful, for instance, to make historical maps more widely accessible. Most static maps, however, offer more than view-only functionality: they may present an interactive view to the user by offering zooming, panning or hyperlinking to other information. The much-used "clickable map" is an example of the latter and is useful as an interface to spatial data. Clicking on geographic objects may lead the user to quantitative data, photographs, sound or video clips or other information sources on the Web.

The user may also interactively determine the contents of the map by choosing the data layers and even the visualization parameters, and by choosing symbols and colours. Dynamic maps are about change; change in one or more of the spatial data components. On the Web, several options to play animations are available. The Web also allows for the fully interactive presentation of 3D models. Virtual Reality Markup Language (VRML), for instance, can be used for this purpose. It stores a true 3D model of objects, not just a series of 3D views.
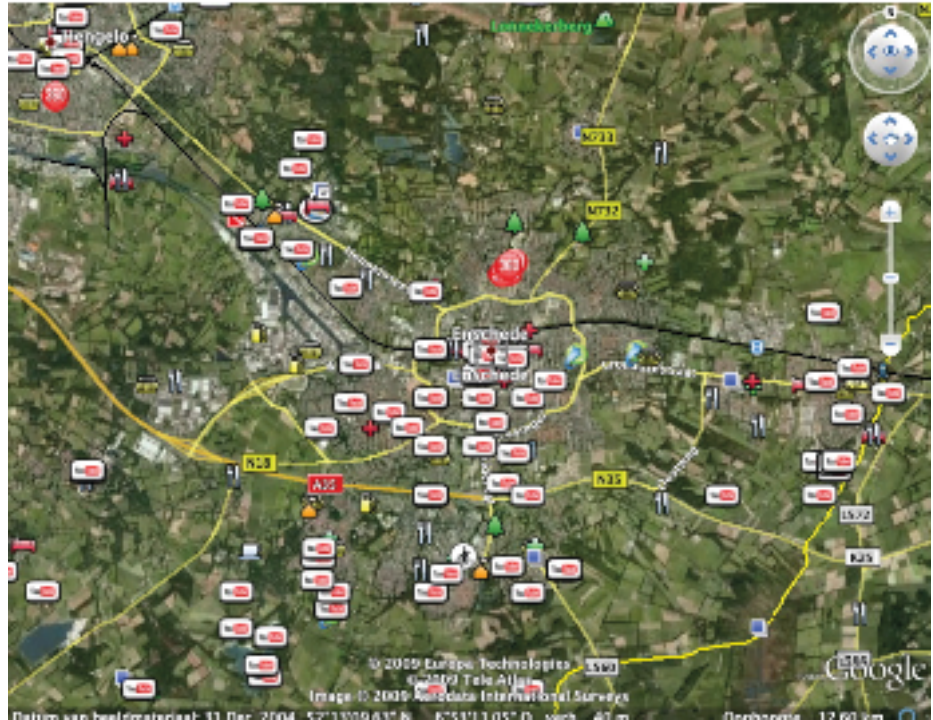
Figure 10.24
Enschede in Google Earth.
Source: [40].

Current mapping options on the Web also allow everyone to contribute their own data. These maps do not always adhere to the guidelines given earlier in this section but they do, nevertheless, have a role to play. Figure 10.24 shows the Enschede area in Google Earth with all available layers switched on. Users can add their own data if they like. Examples of additions would be GPS track and/or waypoints or even 3D buildings. Figure 10.25 is another example of what is known as neo-geography: a detailed view from the Open Street Map of the world (www.openstreetmap.org). All content in this map has been contributed by volunteers. The map shows the area around the ITC building. Note that some work still has to be done on this map, as a road recently built "behind" the building (in fact south of the building, but north of the railway tracks) is still missing.
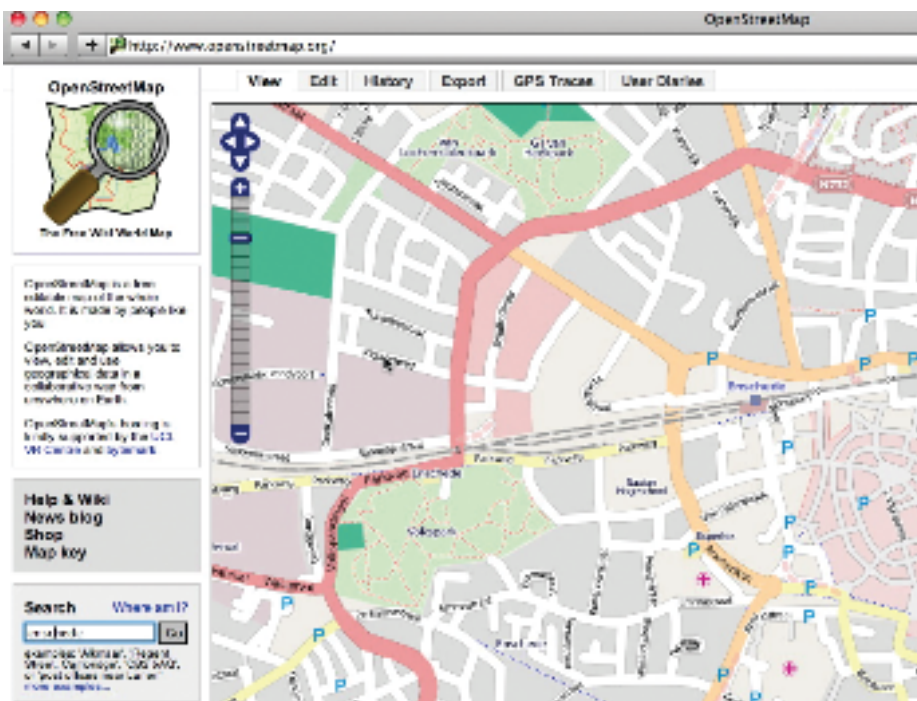
Figure 10.25
Part of the city of Enschede
around the ITC building, as
shown by
OpenStreetMap [87].

Chapter 11

Data integration

Dinand Alkema
Wietske Bijker
Ali Sharifi
Zoltan Vekerdy
Wouter Verhoef

11.1 Introduction

Processes on the planet Earth are complex phenomena that are taking place in space and in time, i.e. in four dimensions. In many of these processes, differences in one dimension (e.g. height above the geoid) can be disregarded, so that two spatial dimensions and the dimension time remain. Despite this simplification, the physical description of the phenomena remains a difficult task. To better understand the processes it often helps if the same geographic region is viewed repeatedly and, if possible, also from different directions and in different wavelength regions. Integration of data from a variety of sources can be a means to retrieving information about processes that would otherwise remain undetected.

Examples of important Earth system processes are:

- Volcanism, earthquakes, plate tectonics

- Ocean currents

- Climate

- Weather

- The living planet (biosphere, agriculture)

- Human activities (urbanization).

Data integration, as the term specifies, concerns the combination and further integration of spatial data. Such a combination may seem at a glance to be simple, but there are various important and challenging issues around it. Indeed, we have seen in earlier chapters that the combination of remote sensing data is not at all trivial. Even if the

images come from geostationary satellites focusing at a single area of land, overlaying these images is still a skilled—although not very difficult—operation. Ground control points can play an essential role in this.

After the overlaying, integration then really concerns the combination of the different wavelength bands, times of observation and, possibly, viewing directions. In addition, the combination of different GIS layers may not be straightforward. Only if GIS layers have the same scale and represent comparable layers of information can overlaying be done sensibly, leading to a possible integration of the information on those layers. An even more challenging case is the combination of a vector layer with an image. For this, the same scale should apply, i.e. a common platform should be decided upon and both sources of data should be made available on this platform so that the overlaying step can be performed. Even then, full integration, i.e. the combination of the information in the two layers, demands that several steps be followed very carefully and close attention must be paid to issues of data quality. Combining layers and images becomes increasingly challenging as the data become more and more complicated. Combining, for example, LIDAR data obtained from an oblique point of view with data from other sources can be done, but usually mathematical transformations have to be applied. Similarly, the combination of data from active sensors with that from passive sensors, of field data with remote sensing data or field data with GIS layers are all activities that require a serious attention throughout.

In the past, much attention was paid to the activities of overlaying and data integration and this is currently still the case. Terms that one typically comes across in this area are conflation, of GIS layers, image fusion, of remote sensing images and registration. Conflation is a somewhat older concept that defines a range of activities that allow one to combine vector layers. Image fusion has had its roots within the ITC research notably by Pohl and van Genderen [92] and is now being further developed throughout the world. Image fusion is the process of combining relevant information from two or more images into a single image. The resulting image will be more informative than any of the input images as it can have complementary spatial and spectral resolution characteristics. In particular the combination of a high spatial and a high spectral resolution image into a single image can be successful. We distinguish fusion at the pixel level, from fusion at the segment level towards fusion at the object level. Image fusion at the pixel level allows the integration of images with each other, but also of an image with other information sources, such as a digital elevation model with images. At the segment level, hence after segmentation of the image, it allows integration with GIS layers. At the object level, hence after a classification, we can also speak about information fusion. Research at this level is still going on. Registration is a term referring to the combination of Earth observation and GIS layers. It is the process of transforming different sets of data into one coordinate system. Data may be multiple images, multiple GIS layers, images from different sensors, from different times, or from different viewpoints. It is used in compiling and analyzing satellite images. Registration is necessary in order to be able to compare or integrate the data obtained from these different measurements and some of it has been dealt with in Chapter 3.

As the focus of this chapter is data integration, it is assumed that the issue of overlaying has been solved, so we can then concentrate our attention on the combination of different layers of information in such a way that new, meaningful information is generated. In this sense, integration is close to modelling, although, as can be seen below, there are some clear distinctions between the two. The *multi* concept will also be introduced as a generic term for the integration of various images at various scales.

Data integration brings with it several issues that play a role in the processing of the data. The most important are:

- data models (raster, vector, TIN, etc.)

- data conversion

- resampling and (dis)aggregation

- gap filling and interpolation

- spectral, angular and temporal effects

- change detection

- visualization techniques

- data assimilation in process models

- multi-sensor approaches.

In this chapter, first a distinction is made between process models and observation models, and it is shown how both types of model can be used together to retrieve more and better information. Next,the *multi* concept in remote sensing is introduced. Data from multiple sources can be integrated to derive more geospatial information of higher quality. Other subjects covered in the chapter are spatial, temporal and spectral scales, and the data conversion issues that arise from data integration. Special attention is paid to change detection techniques, which requires the preparation of the data in similar ways—as is the case for data integration in general.

The chapter concludes with two case studies: one focusing on time series analysis to investigate global climate change; the other looks at the mapping of evapotranspiration in a lake ecosystem by using multiple sources of remotely-sensed data.

11.2   Observation models and process models

To study the relations between object properties and observed spectra, radiative transfer models have been developed. These models enable searching for optimum observation conditions and they can be used for the development of algorithms to retrieve physical properties of observed objects on Earth.

Radiative transfer models that describe the relations between physical and biochemical properties of objects on the one hand, and observed radiation on the other, can be called observation models. In an observation model, the characteristics of the observing instrument, the observational conditions and the observable object's properties all play prominent roles. Characteristics of the instrument and the observational conditions include the viewing direction, the spectral bands used and their spectral and spatial resolution. Object properties are, for instance, canopy LAI (leaf area index) for vegetation and suspended sediment concentration for water. Figure 11.1 illustrates the effect of vegetation canopy LAI on the observed reflectance spectrum in the visible, near infrared and shortwave infrared ranges. These simulations have been carried out with the well-known SAIL model. It demonstrates clearly that for the observation of high LAI values, especially the near infrared part of the spectrum is more sensitive to LAI than the visible part.

In the thermal infrared spectral region, surface temperature and emissivity are important object properties. Examples of other well-known physical theories in which the observation conditions play a central role are, for instance, Einstein's theory of relativity and the theory of quantum mechanics. Planck's law of black-body radiation can

leaf area index
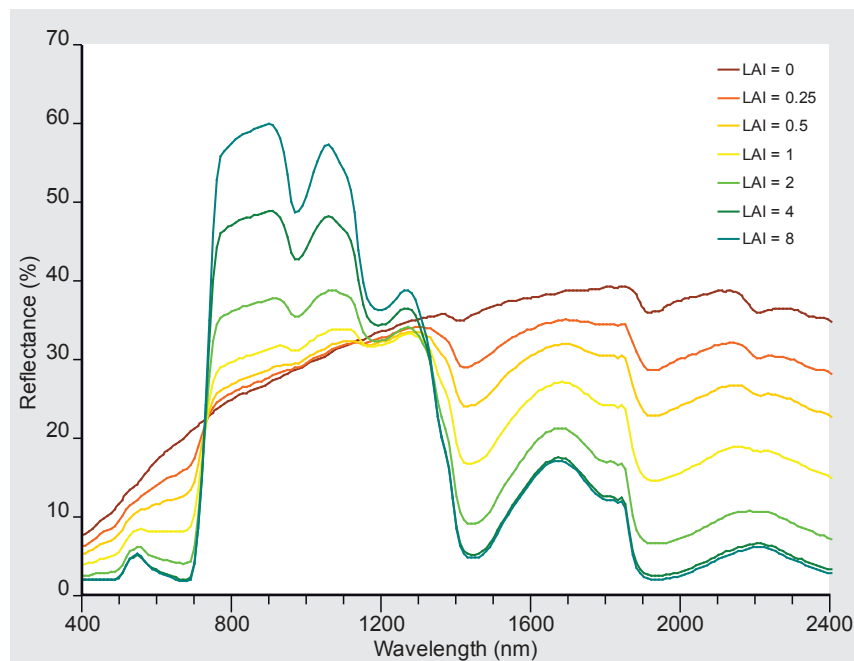
Figure 11.1
Changing vegetation spectra as a function of LAI.

also be considered an observation model as it predicts the observed radiation spectrum as a function of the object's temperature.

Most observation models relevant for remote sensing applications describe scattering and absorption of radiation in various media, such as the atmosphere, water bodies, snow, plant leaves and vegetation canopies. They might also include the scanning mechanism and other observational and instrumental properties, such as viewing direction, spectral and spatial resolutions, and signal-to-noise levels. Observation models may also be coupled. This is very useful, since most remote sensing observations involve a mix of several media, e.g. the combination soil–leaf–canopy and that of sea bottom–water–atmosphere.

On the other hand, we have process models in the Earth sciences that describe the evolution of geo(bio)physical surface properties in time, independently from remote sensing observations. Examples of such process models on various time scales are, for instance, numerical weather prediction models (NWPs), vegetation growth models, hydrological models, oceanographic models and climate models.

Process models in the geosciences usually rely on regular observations at many locations spread over a large area. Traditionally, these observations were mostly made in the field with a variety of instruments. Remote sensing techniques have tremendously increased the capability of spatial sampling and the consistency of the surface parameters measured. RS instruments are mostly sensitive to many physical properties of the surface, some of these may not belong to the set of properties that the user is interested in. Exceptions to this are the mapping of sea-surface temperature, laser altimetry and gravimetry, which are measurements of direct geophysical interest. In the majority of cases, however, there are only indirect relationships between what is observed with the instrument and the physical object properties of interest. In these cases, the use of observation models becomes an attractive option, since these models describe the relationships between all object properties relevant for the observation and the observed remote sensing data.

In general, remote sensing observations can be related to a number of object properties, but also to several other influences, for example atmospheric effects. On the other hand, some surface properties may have no effect at all on any of these observations. All these considerations lead to the following categories of physical quantities within the context of remote sensing:

Category 1 Primary RS observables, i.e. TOA (top-of-atmosphere) radiances;

Category 2 External variables that influence RS observables, e.g. atmospheric variables, Sun angle, view angle;

Category 3 Surface properties not of interest to users but which do have an influence on RS observables (e.g. leaf thickness influences leaf reflectance and transmittance, but will seldom be of interest to a user);

Category 4 Surface properties of interest to users that also have an influence on RS observables (e.g. leaf area index, LAI);

Category 5 Surface properties of interest to users that have no influence on RS observables (e.g. trace pollutants in water are of great interest but they are not detectable).

This is summarized visually in Figure 11.2. For users of remote sensing data, only quantities from category 4 are really of interest. Category 5 is also of interest to users, but undetectable by RS techniques. These quantities can only be measured by other means. Categories 2 and 3 have an impact on the observations, but are not of interest to users. Nevertheless, it is necessary to take them into account, since the observations are sensitive to these factors. Ignoring them might lead to an incorrect interpretation of the observed data.
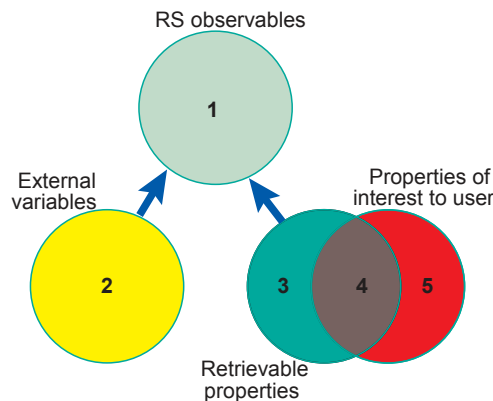
Figure 11.2
Earth Observation variables and their meaning for users. The numbers correspond to the five categories defined in text.

A complete observation model will describe the relations between quantities of categories 2, 3 and 4, on the one hand, and the quantities from category 1 on the other. Most radiative transfer models are, however, not complete, as they describe only radiative transfer in particular medium, such as water, soils, plant leaves, vegetation canopies or the atmosphere. Nevertheless, a complete observation model can be constructed by linking several sub-models together. This is not yet common practice, but is, nevertheless, strongly recommended since conclusions that are based on the interpretation of RS data with only one sub-model may not be reliable.

radiative transfer models

Finally, it should be noted that division over the five categories is not fixed but, rather, depends on the particular remote sensing techniques applied. For instance, object

height has no direct effect on passive RS observations, but it can be measured using active techniques such as laser altimetry. Besides, which properties are of interest to the user is strongly discipline and application dependent.

Observation models and process models can supplement each other to enhance the quality of the interpretation of remote sensing data and to fill gaps in time that occur when observations are not possible owing to clouds or some other cause. Figure 11.3 shows possible interactions of observation models and process models with EO data and existing geographic information (GIS and ground measurements, supplemented with decision-support systems (DSSs)). A central role is played by the GIS database, which provides a common geographic reference. The diagram shows how Earth observation data provide a series of snapshots of the situation on the Earth's surface (green triangle) and how this monitoring of the surface feeds a process model that is updated with actual data (purple triangle). The process model provides information to the decision-support system, which supports management actions aimed at controlling/mitigating the process. A good example of this is a water management system, in which one might decide to allocate water for irrigation if the observed vegetation appears to suffer from drought stress (see case study in Section 11.9).
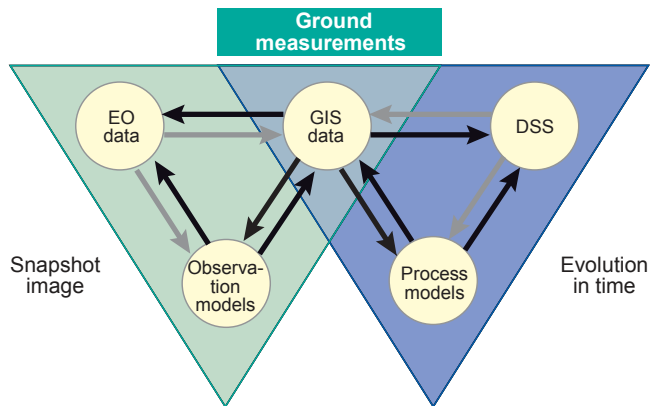
Figure 11.3
Interactions between observation models and process models. Light arrows indicate less likely (but possible) interactions, dark arrows the more obvious ones. The GIS database provides the common geographic reference.

GEOSS

The philosophy of combining many geospatial data sources in order to retrieve more and better information from Earth observation data is expressed in the GEOSS 10 year implementation plan [37], which states:

"Under GEOSS, national, regional, and international policy makers are collectively harmonizing observations, real- or near real-time monitoring, integration of information from *in situ*, airborne, and space-based observations through data assimilation and models."

How the concept of GEOSS could be applied in practice is illustrated in Figure 11.4, which shows a modelling system that simulates images recorded by various sensors on board Earth observation satellites. The heart of the system is a generic RS (observation) model that takes data from a GIS as input and produces as output simulated imagery at the correct spatial resolution and for the spectral bands of the simulated sensor. The RS model includes atmospheric effects and produces top-of-atmosphere (TOA) radiance images for all required spectral bands. The satellite data distributor also provides calibrated TOA radiance data, so this product can be compared to the simulated data.

This comparison is illustrated by the scale symbol, to illustrate the balance between the noise characteristics of the sensor, on one hand, and uncertainty in the surface properties on the other. If simulated and actual satellite images do not correspond
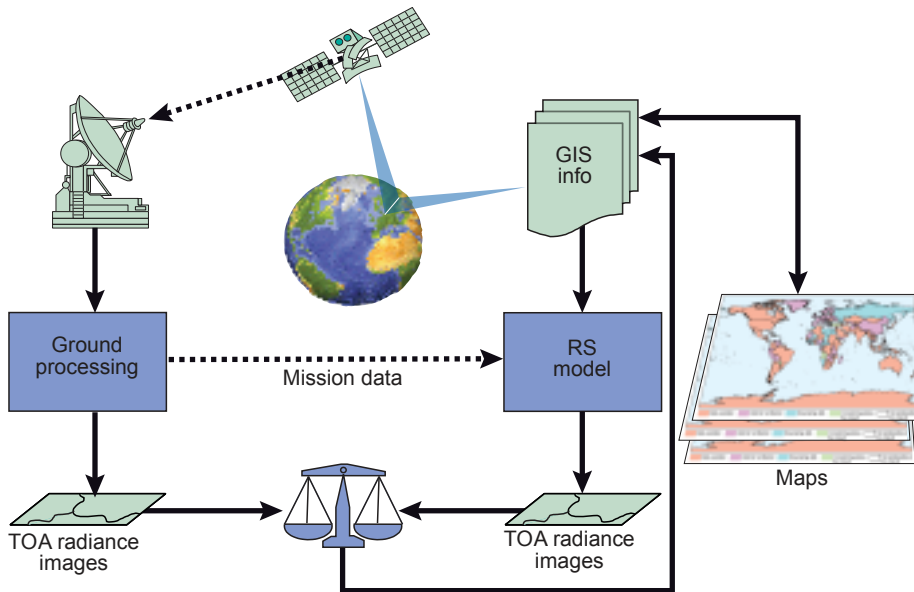
Figure 11.4
A generic image simulation system that produces images that can be compared to actual satellite images. A feedback mechanism minimizes the differences between both images by adjusting the GIS information. Updated GIS data may be transferred to new maps.

sufficiently well, the GIS information is adjusted until the error becomes acceptable. In the GIS, multiple layers of vector and raster data are stored and, in combination with attribute information and values of physical quantities expressing the surface properties, this information is used as input for the radiative transfer sub-models (e.g. for soils, leaves, vegetation canopies and the atmosphere) of the RS model. In other words, the GIS system provides the surface properties as well as their geographic location.

Both spatial data and attribute data (properties) can be in error, and different actions should be taken according to the kind of error. Geometric errors require a correction of the geographic position of one or more objects, whereas errors in surface properties only require the adjustment of these properties. Although the system as sketched is very complex, it has a high degree of flexibility with regard to sensors and geometries, so it would be possible to bridge gaps among the variety of sensor systems that are orbiting the Earth, thereby facilitating the assimilation of data from different sources (as promoted by GEOSS).

11.3   The *multi* concept in remote sensing

Remotely sensed data are often multispectral, sometimes multi-angular, and in some cases also multi-temporal, for instance when time series of satellite data are analysed to discover changes in surface properties or to monitor processes on Earth. If the spectral, angular and temporal domains are exploited to retrieve information about the surface, the data analysis and processing operations become more complicated, but one can retrieve more information from the data. This is why data integration is useful.

A few examples of multiple data observations are:

- colour photography

- multispectral remote sensing

- hyperspectral imaging

- multi-temporal image analysis

- multi-frequency and dual polarization SAR (synthetic aperture radar)

- multi-angular optical observations

- day–night thermal images.

An example of simulated multi-angular observations with a hyperspectral sensor is given in Figure 11.5, which shows how for a sparse vegetation object the observed reflectance spectrum in 201 bands changes with the image acquisitions from space under 7 different directions.
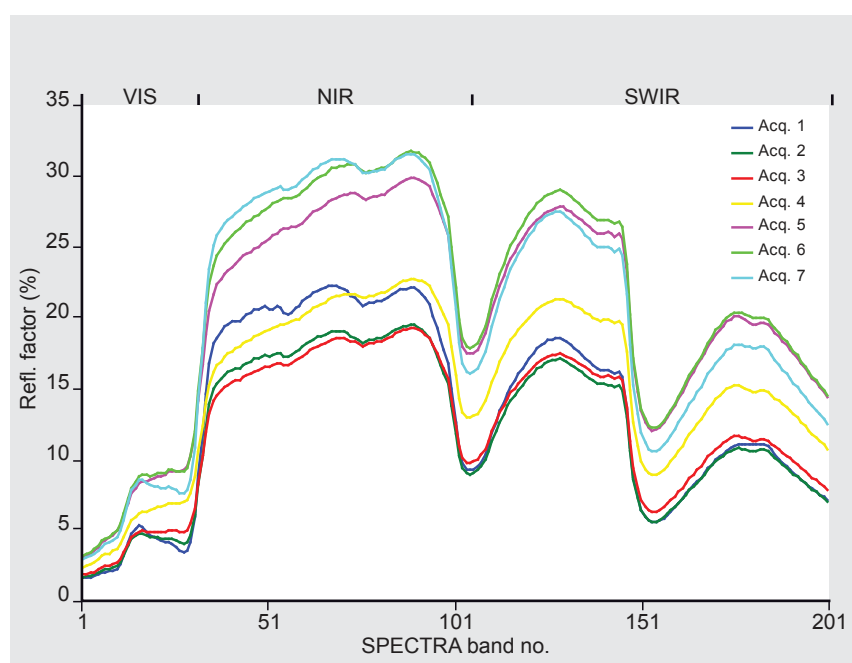
Figure 11.5
Surface spectra under different directions (model simulation)

Remote sensing data can not only be combined with other remote sensing data, but also with existing geospatial information, for instance the geographic information stored in a GIS.

Some examples are:

- digital elevation models (DEMs)

- land use information (GIS data)

- field measurements

- predictions from process models.

If predictions from process models are combined with remotely sensed data, the process model can be continuously updated with new observations. This is called data assimilation, a technique that is intensively applied in operational weather forecasting.

data assimilation

Combination of Earth observation data with other types of geospatial data is highly recommended since existing information can be essential for improving the interpretation of remote sensing data. The automatic classification of agricultural crops from multispectral image data is such an example. Pixel-by-pixel classification usually gives many errors, owing to sensor noise and field heterogeneity. However, if parcel boundaries are known from a GIS database, one can classify all pixels within a field as a group, which reduces the number of misclassifications enormously, provided, of course, that the group as a whole is correctly classified.

Data can be integrated in an almost infinite number of ways. Results from data integration can, again, be combined with other geospatial data to produce yet other new information, and so on. Therefore, only the most obvious forms of data integration will be dealt with in this chapter.

Although data integration can be very useful, there are also some requirements that have to be fulfilled for it to be effective:

- geospatial data have to be accurately co-registered in a common grid;

- time gaps between the various data layers have to be known and accounted for;

- systematic effects due to the atmosphere, the viewing angle, the Sun angle, etc., must be corrected for or taken into account.

In particular, if data from multiple sensor systems are integrated, one has to be aware of differences in their spectral sensitivities, wavelength bands, viewing angles, spatial resolutions, etc. Radiative transfer modelling can be applied to bridge the differences in spectral characteristics and viewing geometry of the various sensors. Other forms of modelling (e.g. 3D object modelling) are sometimes required to aid in the analysis of multi-angular data, for instance to differentiate true changes from apparent changes (e.g. shadows) due to a different viewing direction.

Data integration also comprises the incorporation of non-spatial information or *point* data from field measurements. These data have to be associated with precise moments in time and with precise geographic locations, or with some time interval and fuzzy-defined regions. Thus, here the important issue of the *representativeness* of this information for the associated time interval and geographic area comes into play.

representativeness

In general, data integration forces us to consider the uncertainties or inaccuracies of the various data sources available. In some cases, meta-data may contain information about this. When integrating data for some purpose, one has to apply weights to each of them, so that the final result is a balanced compromise in which inaccurate data receive less weight than those with a high degree of certainty.

The *multi* concept is often applied in remote sensing because multiple observations provide more information, in the sense that objects which look similar in a certain set of observables (spectral bands, times, angles) may look different if the number of observables is increased. A very simple example is the difference between a black & white and a colour photograph. Objects which look similar in black & white photograph may have totally different colours, so they become better distinguishable in a colour photograph. Another example, based on using multi-temporal data, can be found in objects that show similar behaviour as a function of time up until summer, but then start behaving differently. The difference will only become obvious if observations from both "before" and "after" are available.

the multi concept

In principle, multiple observations are always useful, since even if an observation is repeated under exactly the same circumstances with the same sensor (and thus might

seem to be redundant) it is still of use as this helps to reduce the influence of noise. Statistically, the uncertainty of a mean value is inversely proportional to the square root of the number of repeated observations. If the observations are made with multiple sensors, then they provide more information, because different characteristics are measured. For the retrieval of surface properties from Earth observation data it is important to consider which factors determine the retrievability. Of course, only surface properties to which the observations are sensitive can in principle be retrieved, but it is equally important that a change in another surface property does not produce a similar response in the observations. This is the issue of linear dependence. If two surface properties produce a proportionally equal response in all observables (i.e. spectral bands, moments of observation, viewing directions), then there is linear dependence, and one cannot determine which surface property was the cause of the observations. In Figure 11.6, three different situations are illustrated for the case of two surface properties and two observables.

rRetrievability
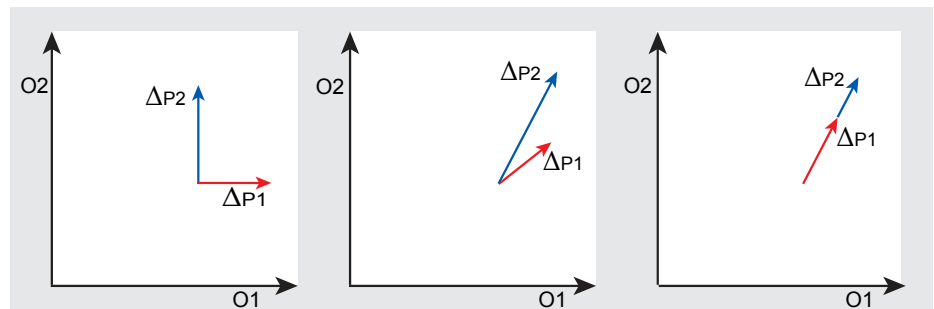
Figure 11.6
Linear dependance: two observables (O) and two surface properties (P).

In the left-hand diagram, property 1 only influences observable 1, and property 2 influences only observable 2. This case is very extreme and seldom occurs in practice, but the conclusion we can draw from the diagram is that in this case property 1 should be estimated from observable 1 and property 2 from observable 2. The middle diagram shows the more common situation that both surface properties influence both observables, albeit in different ways. In this case both properties can still be retrieved mathematically, since we can solve the associated system of two equations with two unknowns. The right-hand diagram illustrates the case of linear dependence. A change in both properties produces a similar change in the observables—perhaps of a different magnitude, but in the same direction. In this case one cannot retrieve both properties separately.

To summarize, the two main conditions for good retrievability are a high sensitivity and linear independence. The chance of encountering cases of linear dependence decreases with the number of independent observations, and since multiple observations are mostly independent, more surface properties can be retrieved and with a higher accuracy.

An inherent part of the analysis, as well as the representation, of multiple observations is visualization. Many visualization techniques have already been explained in Chapter 10, but some additional techniques are shown in Sections 11.7–11.9 to show how linear changes and the dynamics of periodic processes can be visualized.

11.4 Spatial, temporal and spectral scales

In the spatial, temporal and spectral domains that play a role in Earth observation, one can define the concepts of resolution, sampling interval and scale. Resolution describes the ability to resolve small details (in space, time or the electromagnetic spectrum). Sampling interval refers to the distance between two successive observations, while scale describes the total range of observations in a collection of data.

According to the sampling theorem of Nyquist and Shannon, which states that

> "If a function $x(t)$ contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced $1/(2B)$ seconds apart",

the ideal sampling interval should be equal to half the resolution, which means two samples per resolution cell. Otherwise, one loses information (under sampling) or one samples more densely than necessary (oversampling). In practice, however, sampling interval and resolution are often roughly equal, although it is good to keep in mind that resolution and sampling interval are two different things. Note, by the way, that in the above example the time domain was taken as the basis, but in the spatial domain the same considerations apply.

resolution vs. sampling interval

Scales are more applicable to processes in the spatial and temporal domains than to those in the spectral domain. For a number of important Earth system processes, their corresponding scales are roughly indicated in Figure 11.7, which has logarithmic X- and Y-axes in order to accommodate the large ranges to be considered in the spatial and temporal domains. In the spatial domain this goes from 1 mm to 40,000 km (the circumference of the Earth), and in the temporal domain from one second to a century.

Figure 11.7
Temporal and spatial scales of some Earth system processes.

Spatial and temporal scales can not only be attached to processes, but also to observations. An example is given in Table 11.1, which summarizes the spatial and temporal scales of a few well-known Earth observation systems.

As well as spatial, temporal and spectral scales, which are more related to the act of making observations, it is equally important to consider different levels of spatial

Table 11.1
Scales of RS observations.

Sensor	Spatial scale	Temporal scale
Meteosat	Hemisphere	15 minutes
NOAA-AVHRR	3000 km	daily
Landsat TM	180 km	16 days
SPOT	60 km	26 days
		(pointable)

levels of spatial aggregation

aggregation in various Earth science processes. These aggregation levels are mostly ordered hierarchically. For instance, a forest consists of trees, which have trunks, branches and twigs, on which one finds leaves or needles, and so on. Similarly, an agricultural area may be described in terms of parcels, lots, cropping fields, individual plants, stems and single leaves.

Analysing multiple layers of geospatial data in a meaningful and coherent way requires the co-registration of all these layers to a common spatial grid or reference. The common grid spacing chosen is application-dependent and several considerations may play a role in that choice. In some cases the preservation of high levels of spatial detail is most important, and in other cases high levels of radiometric precision may be more relevant. In the first instance, one would probably choose a common grid spacing that accommodates the data layer with the highest spatial resolution, while in the second instance one may choose a grid with a wider spacing. Figure 11.8 illustrates the display of two images with different grid spacing and orientation for a part of Enschede. The images have the same georeference, yet the pixel size and orientation are different

resampling and aggregation

A common grid also requires a *resampling* of those layers that have a different spacing and/or orientation. In some cases, especially when an image of high spatial resolution is converted to a less dense grid, the resampling has to be combined with *aggregation*, e.g. implemented as a low-pass filter applied to the input layer so as to exploit the high spatial density of that layer, in order to increase the radiometric accuracy and thus reduce noise.

In addition to spatial resampling and aggregation, other operations sometimes have to be applied to the data to condition them for data integration. This would be necessary when, for example, analysing a long time series of satellite data for which the calibration data have gradually changed in the course of time—or even suddenly, for example owing to the replacement of an existing satellite with a new one. In such cases, the calibration data have to be corrected in order to obtain a time series that is free of these artefacts.

Sun-synchronous

In time-series analyses of NOAA AVHRR data, however, another gradual effect was observed that could not be corrected. This effect is known as the orbital drift problem. Each NOAA satellite has a Sun-synchronous orbit that is not very precise, which means that after a number of years the moment in time at which the Equator is crossed gradually increases by a couple of hours. See Figure 11.9, which illustrates this for several satellites from the NOAA series. This means that observations coming from a satellite that is at the end of its life are not comparable to similar observations made at the beginning of its life. Especially in the Southern Hemisphere, this led to local solar times of observation very late in the afternoon. However, effects caused by drastically changing illumination conditions cannot be corrected, so in such cases one has to reject data if conditions deviate too much from normal.

If the calibration data of one or more satellites have changed, we can sometimes compensate for these effects by using a target on Earth that has stable reflectance proper-
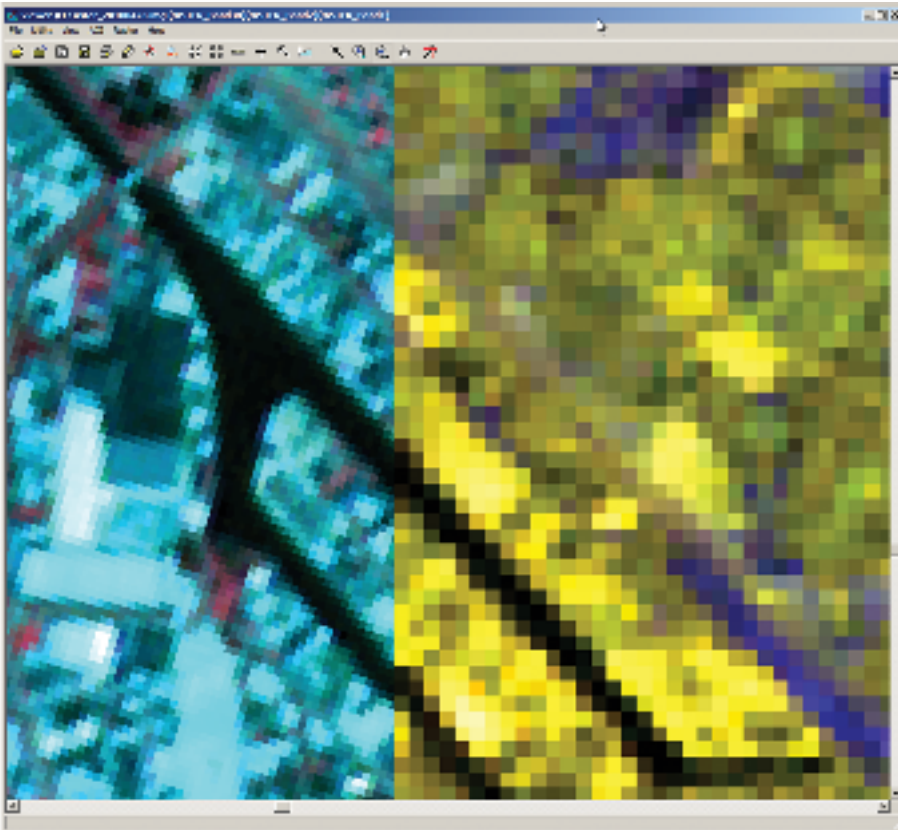
Figure 11.8
Difference in grid spacing and
orientation.
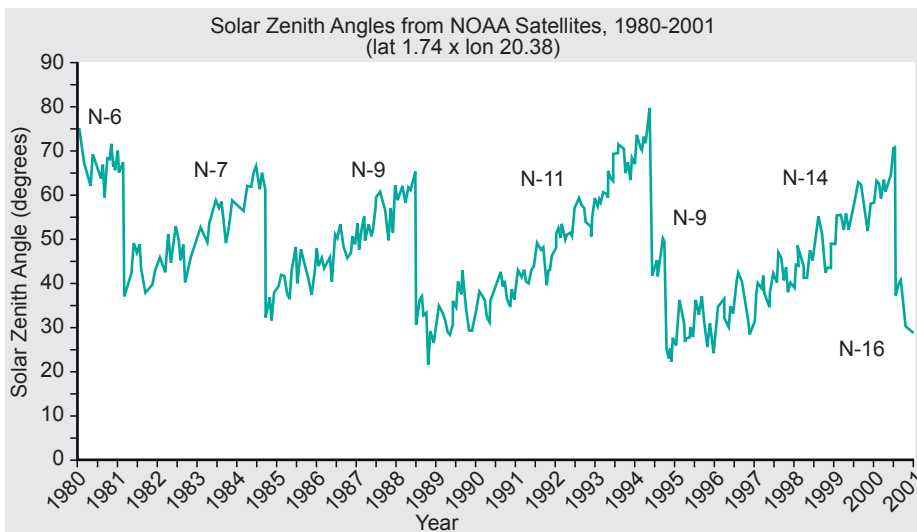
Figure 11.9
Course of the local solar time
of observation near the
Equator from 1980 to 2001
for the NOAA series of
satellites. Source `http:
//classic.nerc.ac.uk/`

ties. In one such case, the Libyan desert in Africa was used as a stable reference target with a constant NDVI value, and all data were recalibrated so that the same NDVI resulted for that area during the whole time series of 18 years.

If two sensors with different spectral characteristics are involved in the data integra-

tion, then one can try to match the spectral bands of one sensor with the closest bands of the other sensor, but one has to be careful in doing so, since subtle spectral differences between objects on the ground will be observed differently by both sensors, so this cannot be completely corrected. Figure 11.10 shows the spectral response functions for the ASTER and MODIS sensors on board the Terra satellite, together with some spectra of surface reflectance. This clearly illustrates the problem of combining different sensors in spectral regions where surface reflectance is rather variable, such as in the so-called red-edge region (around 700 nm).
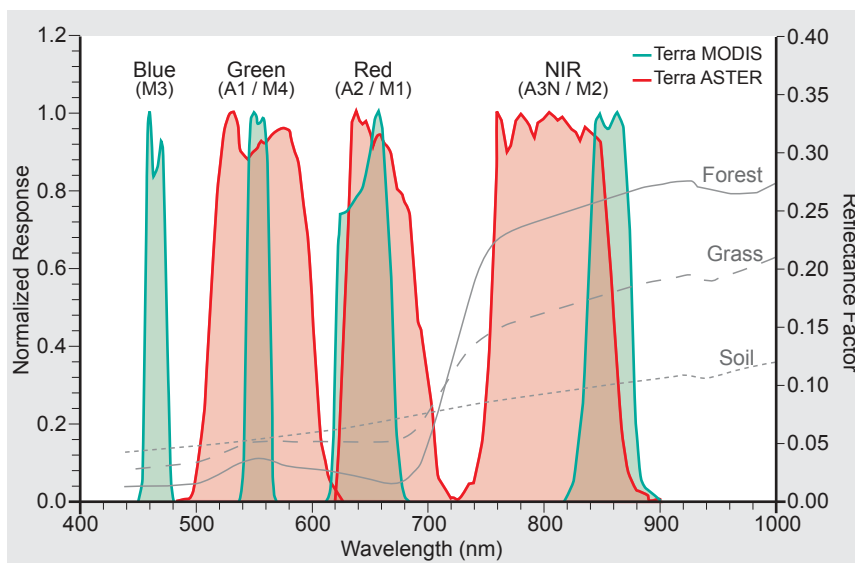
Figure 11.10
Spectral response functions for the ASTER and MODIS sensors on board of the Terra satellite and reflectance of three surface types.
Source: [75].

Similar considerations apply in the case of multiple observations from different satellites that are treated as synoptic observations of one moment, while actually there may be a time difference of days or a local time difference of hours. These cases are very difficult to handle correctly, but radiative transfer modelling can be applied to bridge the difference in the Sun–target–sensor geometry and process models might be used to account for time differences.

Apart from the more sensor-based data-conversion issues already mentioned, sometimes conversion operations are also necessary to bridge differences in data formats related, for example, to:

- computer data types (byte, integer, float)

- image data organization types (raster BIL, BSQ and BIP formats)

- image data formats (JPEG, GIF, BMP, GeoTIFF, HDF, etc.)

- raster <> vector (points, lines, polygons)

- vector TIN to contour lines

- physical unit conversion.

11.5 Data integration issues in GISs

Integrating data sets in a GIS often results in an improved understanding of the problem/phenomenon at hand. One could even say that data integration is the raison d'être of GISs; in any case, data integration certainly facilitates further analysis of the data.

In real-life projects the user often has to integrate data by:

- merging mismatched data layers

- choosing, in cases where two data sets of the same features exist, which set should be preferred (based on criteria that need to be defined);

- solving, for example, problems such as changes in administrative units (merging or splitting of areas) and matching these with data that only refer to an administrative name or code.

Moreover, there is always a need to merge non-spatial (statistical data, social behaviour data, . . .) with spatial data. With volunteered geodata and with crowdsourcing (Web 2.0), data integration becomes both more tricky and also more important. In this respect, meta-data and lineage documentation are essential for proper data integration. The merging of mismatching data layers might require dealing with:

- mismatchings in area (spatial extent)

- mismatchings in level of detail (scale)

- mismatchings in projection (georeferencing)

- mismatchings in time

- mismatchings in accuracy

- mismatchings in data format/type (tabular and spatial data)

- mismatchings in purpose for which the data are being collected.

11.6 Change detection

Change detection is a particular application of Earth observation in which data integration is required before one can concentrate on the observable changes. This application is susceptible to the problem that changes in the observation system can be confused with the actual changes in the target objects themselves. Once again, in this case radiative transfer and observation modelling, adapted to the sensors at hand, may help to separate apparent changes from true changes. Here, true changes are changes in the object properties, while apparent changes are only related to changes in the observation conditions.

true and apparent changes

In change detection, there are different kinds of change to be distinguished:

- *Gradual changes* (sometimes called linear changes) are involved in climate change investigations, deforestation, urbanization, etc.

- *Sudden changes* are mostly related to disasters (floods, earthquakes, volcanic eruptions, fires, etc.)

- *Periodic changes* are related to the daily and yearly cycles of solar illumination and warming, giving rise to the diurnal cycles of daylight and temperature and the seasonal patterns of vegetation growth.

This section will focus on how data integration is applied in change detection. We will focus on changes on the surface of the Earth, but there is certainly an overlap with the techniques used in other image processing sciences, like medical imaging.

Changes are caused by processes. These processes can be natural, man-made, seasonal, deterministic or random. To determine appropriate data sources for detecting change, several characteristics of the process need to be known, such as speed, duration, observables, area coverage and seasonality. Section 4.7.1 (and Section 11.1 and Figure 11.2) explains how these characteristics can be used to determine spatial and temporal coverage, time and frequency of observation, spatial resolution or scale, and observables. The expected size of the change determines the degree of sensitivity needed in the analysis.

Change detection can be carried out at various levels of detail or sophistication, depending on the interests of the user. This may include answering some or all of the following questions:

- Has there been a change (detection)?

- What is the nature or type of the change (identification)?

- What is the area covered by the change (area)?

- What is the spatio-temporal pattern of the change?

Over the years, several categories of change detection techniques have been developed. These techniques will be discussed in the remainder of this subsection, focusing on whether they can be used for detection, identification, area and/or spatio-temporal patterns of change. The definition and description of categories is mainly based on [70].

Algebra techniques for change detection

This group of techniques includes all kinds of algorithms that are based on combinations of values of a pixel in subsequent images, such as image differencing, image ratioing, vegetation index differencing, image regression and change vector analysis.

image differencing

Image differencing is a band-by-band, pixel-by-pixel subtraction of two images whereby the resulting change image has the same number of bands as the input images. Each band of the change image contains the differences between the spectral values of the pixels in the two original bands.

image ratioing

Image ratioing is a band-by-band, pixel-by-pixel ratio of two images whereby the resulting change image has the same number of bands as the input images. Each band of the change image contains the ratios of the spectral values of the pixels in the two original bands.

vegetation index differencing

For vegetation index differencing, vegetation index is calculated for each image; the change image contains the differences between the vegetation indices for each pixel.

image regression

For image regression, a relation is established through regression between two images of different dates. The relation is used to predict pixel values in the second image, which are then subtracted from the first image. Regression reduces the effects of sensor, atmospheric and environmental differences between the two images. Development of suitable regression functions can be difficult.

In change vector analysis a spectral change vector is calculated, which describes the direction and magnitude of the change between two dates and a total change magnitude for each pixel. The total change equals the Euclidean distance between end points in an *n*-dimensional change space; any number of spectral bands can be processed in this way. The method produces detailed change information.

change vector analysis

All methods in the "algebra" category rely on the selection of a threshold on the change image, to separate noise and apparent changes from true changes and to determine the change areas. The choice of the threshold is often difficult and arbitrary. With exception of change vector analysis, these methods cannot provide a complete change matrix, so there is no complete identification of the nature of all changes. Algebra is often used when the focus is on detecting a very specific change, for example detection of forest fires, where changes in the thermal bands indicate a rise in temperature of the land surface and a threshold determines whether or not there is a fire. Other common applications include deforestation mapping and detection of vegetation change.

change matrix

Classification-based change detection

The techniques in this category all involve some kind of classification of separate or combined images. Some of the most common techniques are discussed in the remainder of this subsection. For a more exhaustive list, see [70].

In post-classification comparison, the images are classified separately and classifications at different dates are compared. The advantages of this technique are the minimization of atmospheric influences and the fact that it generates a complete change matrix. The disadvantages are that sufficient training data are needed for each classification (date) and that there might be systematic differences between both classifications.

post-classification comparison

For spectral-temporal combined analysis, all images are stacked in one data set and classified together, all at once, similar to a multispectral classification with many bands. Changes are identified and labelled. This can be time-saving, but it may be difficult to identify and label the change classes. A complete change matrix cannot be provided.

spectral-temporal analysis

Unsupervised change detection labels spectrally-similar groups and clusters at Date 1, followed by spectrally similar groups at Date 2, and then detects changes. Because an unsupervised algorithm is used, the process can be automated, but labelling the changed areas is not always straightforward, especially in the case of processes (i.e. a series of changes that are part of a process, such as conversion of forest via burnt areas to crops and, finally, pasture).

unsupervised change detection

Hybrid change detection first isolates changed pixels to construct a binary change mask. The change mask then sieves out the changed themes from land use/land cover (LULC) maps.

hybrid change detection

Visual analysis for change detection

The human eye is still one of the most powerful instruments available for detecting and interpreting change. Visual interpretation can be aided by different ways of displaying time series, such as multi-temporal colour composites and animations. Results depend very much on the skills of the analyst and their familiarity with the area.

GIS

Change can also be detected by combining maps and images. GIS overlays on image data can provide a means for better interpretation and for detection of changes, e.g. new buildings, changes in parcel boundaries or forest limits. The results can be directly combined with other data in the GIS, for example to update LULC information. The difficulties lie in reconciling the accuracy of different sources and their registra-

GIS overlays

tion.

Past and current maps of land use/land cover can also be combined and, if necessary, integrated with topographic and geological data. In addition to difficulties arising from differences in accuracy of the sources and their registration, thematic categories in the maps may not always match.

Other methods

transformations

Transformations are essentially data reduction techniques, which reduce redundancy between bands or images and highlight any differences. They have been applied to change detection, but the disadvantage of such transformations is that they cannot produce detailed change matrices, they require (arbitrary) thresholds, and it is difficult to interpret and label the change information they generate.

model-based methods

A number of model-based methods have been developed for very specific purposes. These include the modelling of reflectance of certain classes or for the retrieval of biophysical parameters from spectral characteristics.

object-based change detection

Furthermore, new developments in image classification also result in new approaches to change detection. Image classification based on objects (object-oriented analysis) is employed in change detection. Fuzzy classification allows for the fact that boundaries can be vague or gradual and developments in this field also lead to the development of change detection methods that can adequately deal with changes in gradual boundaries or in objects with vague boundaries. This approach is, however, beyond the scope of the Core module.

The case study on climate change described in Section 11.7 is an example of how one can analyse a mixture of periodic phenomena (annual vegetation growth) and gradual changes (year-to-year variations).

11.7 Case study: Climate change

To find evidence for the impact of climate change on processes on Earth, time series of satellite images are particularly suitable, since they constitute regular observations over a long period of time under comparable conditions. Many processes on Earth are controlled by the yearly seasonal cycle and can, therefore, be roughly modelled as periodic functions, although these processes are also gradually evolving owing to climate change. In order to find evidence for climate change, long-term trends have to be separated from seasonal effects. This can be done by estimating the periodic parameters of each year and by analysing trends in those annual parameters. In this case study, a time series of NOAA NDVI images of the world was analysed to discover trends in global vegetation-growth patterns over the period 1982–2000. In most years, 36 images of the 10-daily maximum NDVI were used as input.

The maximum value compositing (MVC) technique is often applied to time series of NDVI images in order to reduce the effects of cloud cover, atmospheric conditions and large viewing angles. During a 10-day period the NOAA satellite views a given location on Earth every day at approximately the same local solar time, but the viewing angle and the atmospheric conditions may vary a lot during this period. As the NDVI is always *decreased* under cloudy or hazy conditions, as well as for large viewing angles, by taking the maximum NDVI of the 10-day period one hopes that the best observation has been selected. This method removes most cases of cloud cover in most regions, but when the NDVI of a certain location is plotted as a function of time, one still sees some residual effects of cloud and haze.

maximum value compositing

The HANTS (harmonic analysis of time series) algorithm works similarly to MVC by removing negative outliers, not within 10-day intervals but over the whole year. In this case, the signal is modelled by a series of sine and cosine functions that span the whole period and have frequencies corresponding to the base period (e.g. one year) and a number of higher harmonics (waves of higher frequency that fit in the base period). As illustrated in Figure 11.11, negative outliers are removed and a gentle curve is fitted to the remaining points.
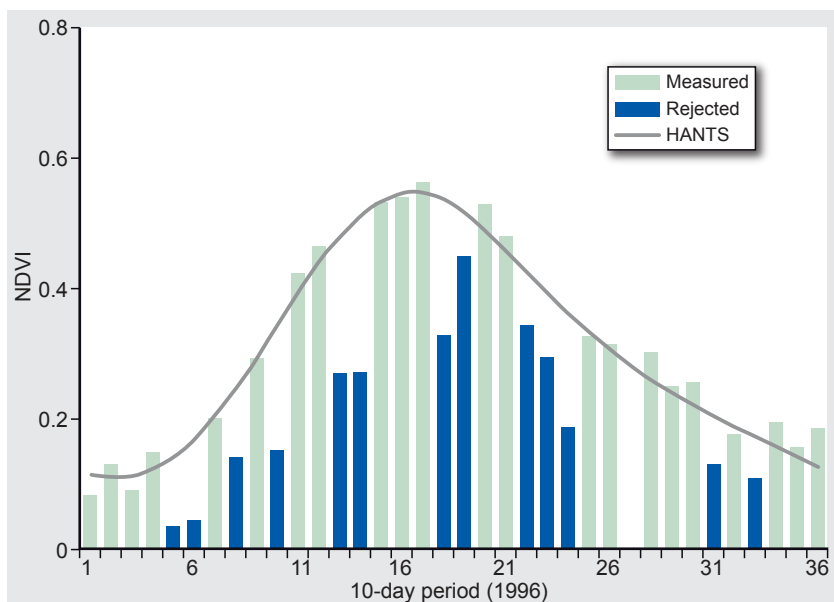
Figure 11.11
Outlier removal and curve fitting using the HANTS algorithm.

Provided a phase is assigned to each frequency, the series of sine and cosine functions can also be expressed as a series of only cosine functions:

$$y(t) = a_0 + \sum_{i=1}^{n} a_i \cos (\omega_i t - \varphi_i), \qquad (11.1)$$

where $a_0 \ldots a_n$ are the amplitudes and $\varphi_1 \ldots \varphi_n$ the phases. The circular frequencies $\omega_1, \ldots, \omega_n$ are chosen in such a way that $\omega_i = \frac{2\pi}{T} i$, where T is the length of the base period. The amplitude of the zero frequency, a_0, plays a special role, since it equals the mean value of the modelled time series. To model vegetation dynamics throughout the year, three frequencies above the zero frequency are usually sufficient. This also means a considerable data reduction, since 36 original data points are then represented by only seven components, namely the mean and, for the three frequencies, the amplitude and the phase.

Modelling with a higher number of frequencies allows more details of the original series to be preserved, but then there is a chance that effects due to cloud and haze are preserved as well. Vegetation growth, on the other hand, is usually a fairly gradual phenomenon and three frequencies are sufficient to follow the most rapid growth spurts. Figure 11.12 visualizes the results for one year (1995) of data for the world.

This kind of visualization reveals that the frequencies 2 and 3 (periods of 6 and 4 months, respectively) are already quite noisy owing to residual cloud cover and haze effects. The mean, the yearly amplitude and the yearly phase look most reliable.

The phase is a number ranging from 0 to $360°$ or, if expressed in radians, from 0 to 2π. Showing the phase as an image in black & white creates a problem, since $0°$ would be shown as black and $360°$ as white, while actually $0°$ and $360°$ represent the same angle. In this case one can use a colour look-up table that is circular in the RGB values, meaning that $0°$ and $360°$ are represented by the same colour. A *rainbow* look-up table (LUT) can be constructed in such a way that it follows the colour sequence blue-cyan-green-yellow-red-magenta-blue, so that the start and the end have the same colour. Such a LUT was applied to the phase images of Figure 11.12.

Phase information is independent of amplitude, but in practice one would like phase information to be considered less important for small amplitudes. The following set of equations provide a way of accomplishing a colour transformation that is controlled not only by the phase but also by the amplitude:

$$R = 127 \times \left[1 + \frac{A}{A_{max}} \cos(P - 240) \right]$$

$$G = 127 \times \left[1 + \frac{A}{A_{max}} \cos(P - 120) \right]$$

$$B = 127 \times \left[1 + \frac{A}{A_{max}} \cos(P) \right]$$

Here P is the phase in degrees, A the amplitude, and R, G and B are the amounts of red, green and blue on a scale from 0 to 255. The constant A_{max} is the maximum expected amplitude. If the amplitude equals the maximum amplitude, the RGB values vary from zero to 254, so a large colour saturation is obtained. This is also demonstrated by Figure 11.13, which shows the curves for R, G and B as a function of P if $A = A_{max}$.

For zero amplitude R, G and B all become equal to 127, so one will observe mid-grey.
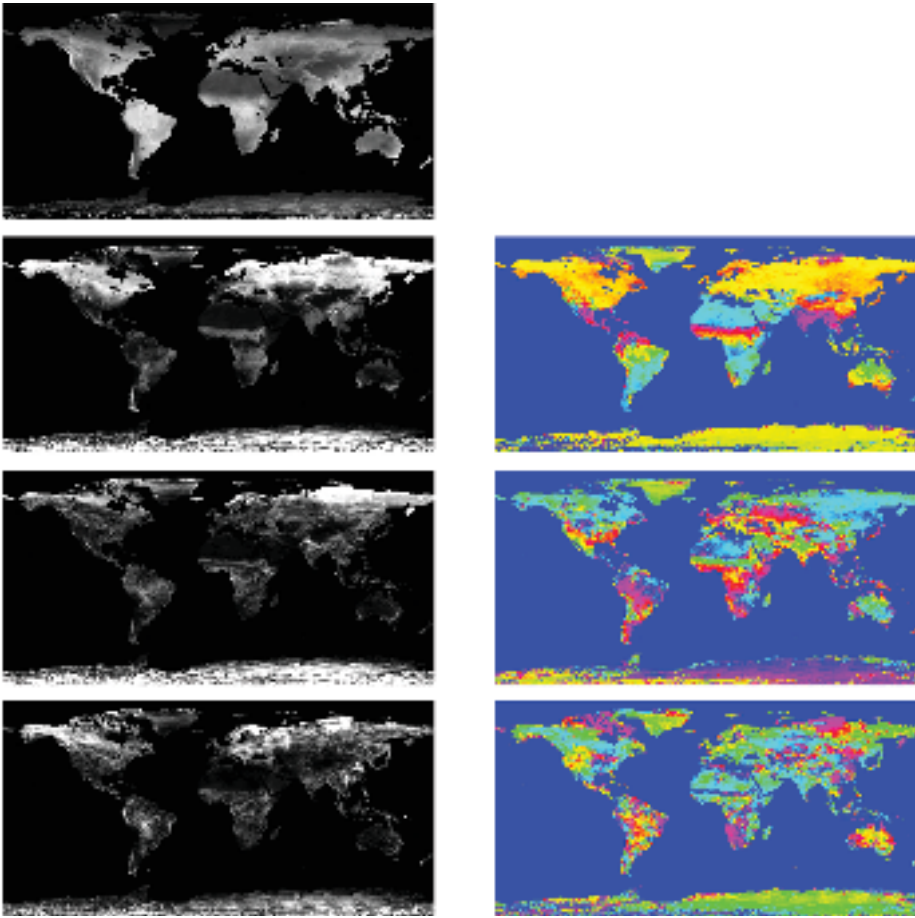
Figure 11.12
Global NDVI dynamics for the year 1995. Amplitudes are in the left-hand column; phases are in shown in the right-hand column in rainbow colours. Frequencies (top to bottom) = 0, 1, 2, 3.
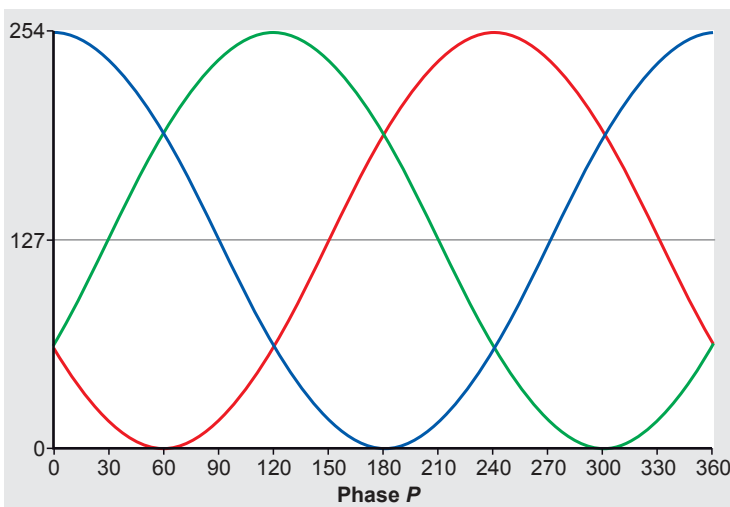
Figure 11.13
Variation of R, G and B as a function of the phase angle P.

It is also possible to include mean NDVI by coupling it to the pixel's intensity. In this case the factor 127 in the above formulas is replaced by

$$\frac{M - M_{\min}}{M_{\max} - M_{\min}} \times 255,$$

where M is the annual mean NDVI. This method of visualization was applied to global data: Figure 11.14 shows global NDVI dynamics for all years from 1982 to 2000.

Figure 11.14
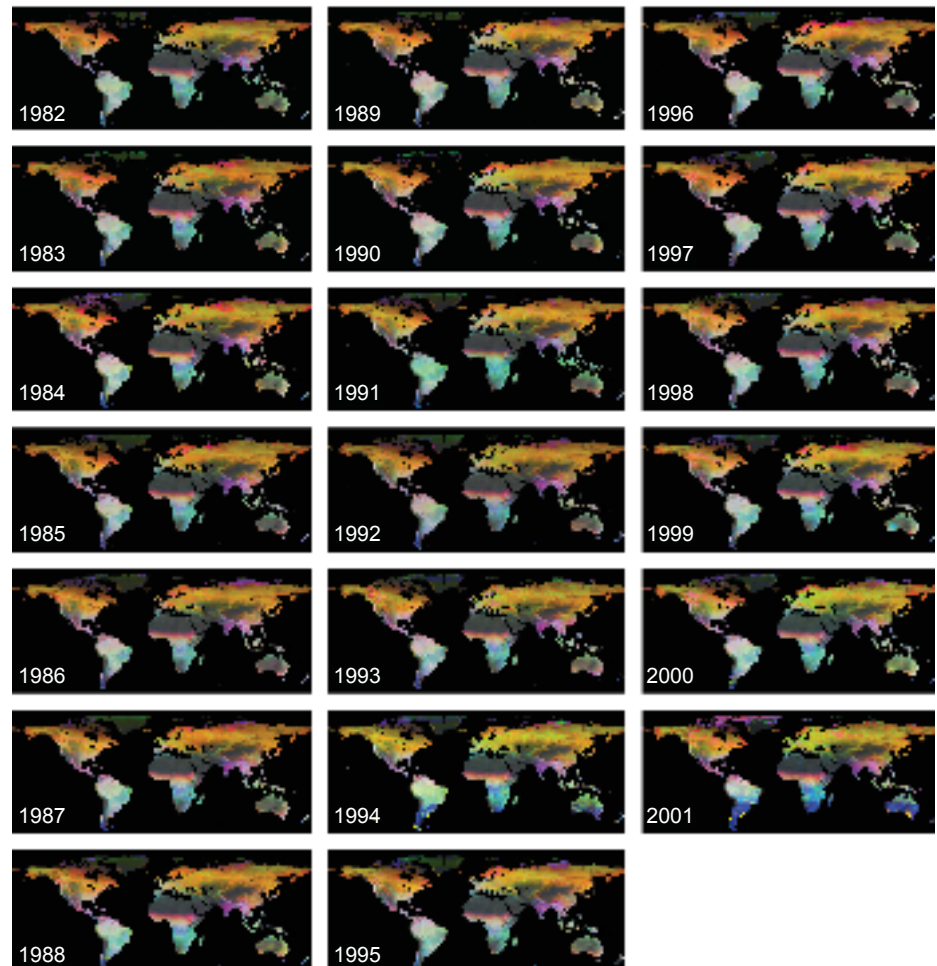IHS representation of global yearly vegetation dynamics for the years 1982–2001. Note anomalous behaviour in southern areas in the years 1984, 1988, 1993+1994 and 2000+2001 owing to the problem of orbital drift.

In order to obtain evidence of climate change, trends in the changes of mean NDVI, its yearly amplitude and its phase were analysed by correlating these quantities with time. In addition, the start and the end of the growing season were established from the yearly growth patterns by finding the intersection of the growth curve with fixed NDVI levels. Next, simple linear regression analysis was applied for each pixel to find the rate of change (i.e. the slope of the regression line) at each location. To display the results, a colour look-up table was used in which negative slopes were shown in blue, positive slopes in red, and zero slopes in white; the degree of colour saturation was used to show the magnitude of the slope.

The result is shown in Figure 11.15, which indicates that trends in vegetation dynamics have taken place all over the world in this period of 20 years. These trends are characterized by a slight increase in annual mean NDVI, increasing as well as decreasing yearly amplitudes, an advancement of the yearly phase, an advanced start

of the growing season, and a lengthening of the growing season. These phenomena are due to increasing global temperatures, especially in winter; CO_2 fertilization; and advances in agrotechnology.

Note that in Antarctica strong trends are also found. However, these have to be considered as unreliable as the associated correlations were not significant. If we only take into account the locations where the correlation coefficient is significant at the 5% confidence level (for 20 observation this means a minimum absolute correlation of 0.45), then the picture shown in Figure 11.16 is obtained.

One can observe in Figure 11.16 that that significant trends are especially clear for India and the Sahel region in Africa and that the clearest trends are indicated by the increasing mean NDVI and the longer growing season in these regions.

These results are based on 20 years of NOAA-AVHRR data, which actually comprised 36 maximum-value composites of the 365 daily NDVI images per year. The NDVI is, in turn, based on the red and near-infrared spectral channels, and the original ground resolution of these GAC (global area coverage) data was 4 km at nadir. These data were aggregated by the satellite data distributor into $1° \times 1°$ geographic cells before their dissemination to users. Roughly estimated, $20 \times 365.25 = 7305$ satellite images in two channels have contributed to the final result (indicating climate change) for each single $1° \times 1°$ pixel, so this represents an extreme example of data integration.

The visualization technique presented in Figure 11.14 for the whole world has been applied to smaller regions as well, using SPOT-VGT NDVI images as input. Only data from the year 2002 were processed to show the influence of topography on vegetation-growth dynamics. In order to express the terrain topography in the image, a technique called hillshading was applied. Here, data from a digital elevation model (DEM) were integrated with the NDVI dynamics by modulating them with the resulting hillshading, as shown in Figure 11.17; the colour circle indicates which colour hue is to be associated with the month of maximum NDVI.
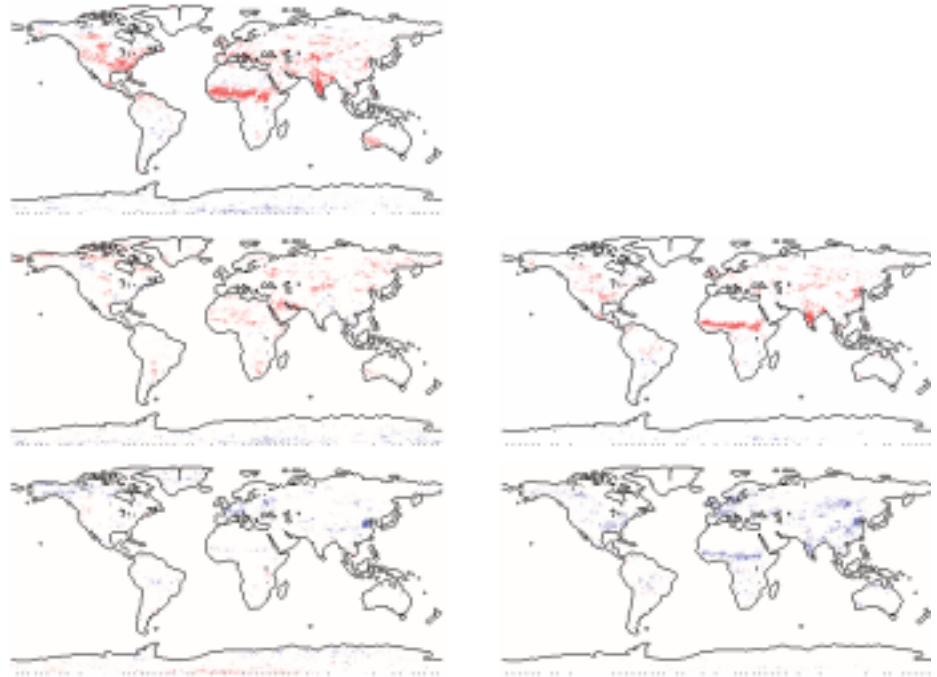
Figure 11.16
Significant (5% confidence level) correlation coefficients (absolute value > 0.45) for trends corresponding to those shown in Figure 11.15. Positive trends are shown in red, negative trends in blue.
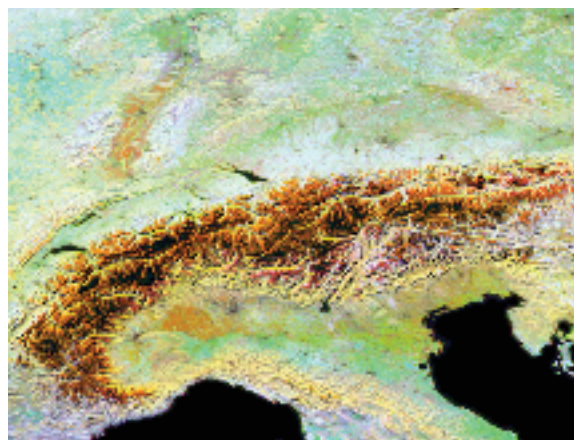
Figure 11.17
Vegetation dynamics in 2002 of the Alps and surroundings. The colour circle shows the relation between the month of maximum NDVI and the colour hue.
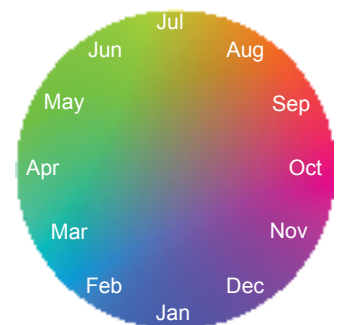
Month of maximum green vegetation (NDVI) and related colour

In this image, one can observe a strong relation between topography and vegetation growth. The Alps show red and brown colours, indicating maximum NDVI in August–September. This is because in mountainous regions there is snow in winter and spring, which delays growth. In very high regions, snow cover is permanent. These have no vegetation at all during the whole year, so they appear black. In the image one can also clearly recognize the Upper Rhine Valley (orange), the Black Forest (white), the Po Valley and the Apennines in Italy, and a large area west of Milan where rice is grown (orange). The grasslands to the north of The Alps are white, indicating permanent green vegetation cover throughout the year.

The same SPOT-VGT data for the whole of Europe were used to make a cloud-free, synthesized time series of images that show winter snowfall and its withdrawal in spring and summer. In this case the original SPOT bands in the red, near infrared and

shortwave infrared were used, and positive outliers were removed using the HANTS algorithm to obtain cloud-free data. Snow has a high reflectance in the visible and the near infrared, but it absorbs radiation in the shortwave infrared. In an RGB composite of NIR, SWIR and RED (in that order), snow appears in purple (magenta), since the SWIR component is missing, and only NIR and RED remain, which control red and blue, respectively, so we obtain magenta colours. Such a sequence of images can be played as an animation, which is an alternative way of observing the processes of vegetation growth and snow cover. The individual images are shown in Figure 11.18.
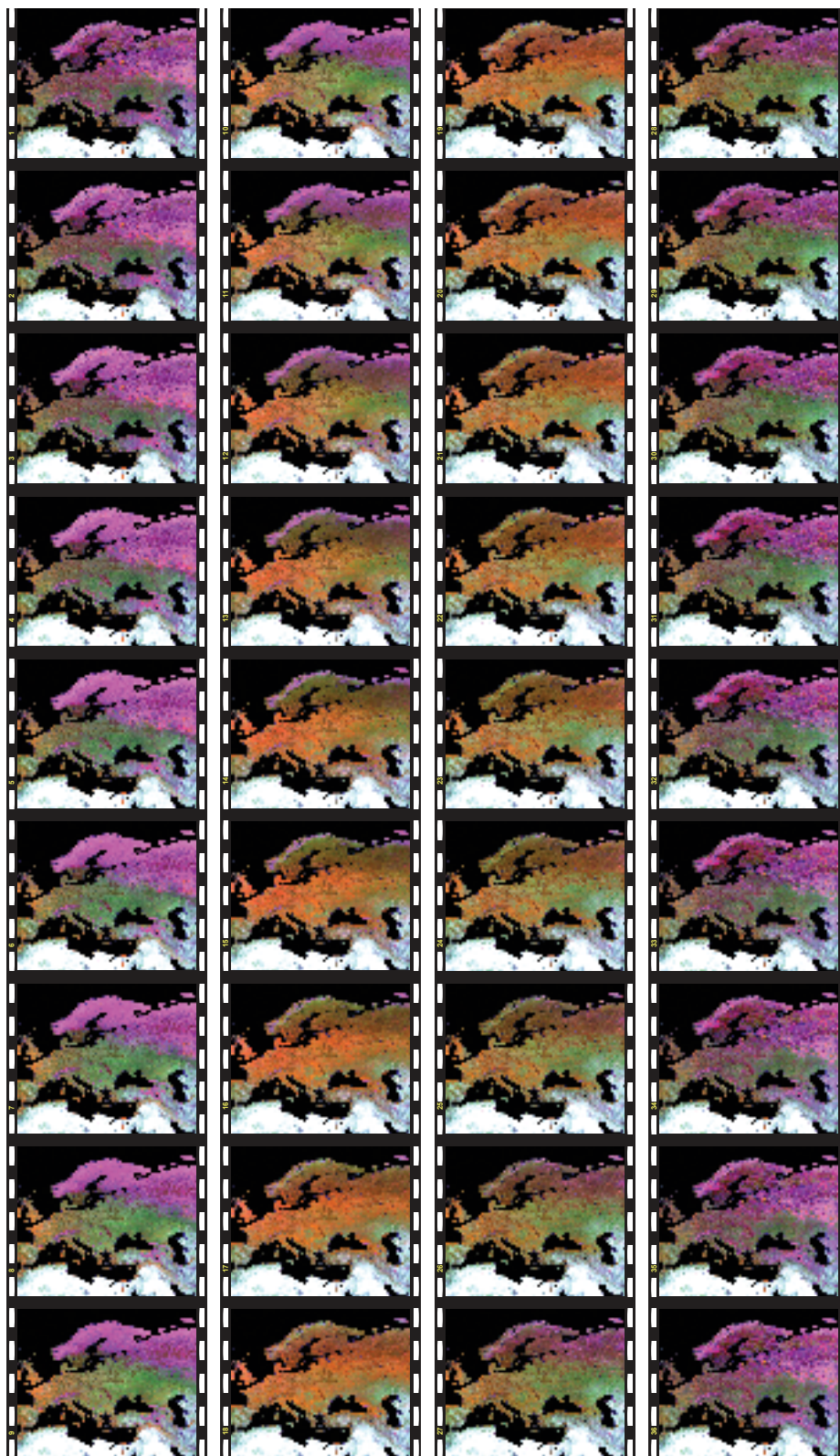
Figure 11.18
Synthesized cloud-free time series of 36 SPOT-VGT images of the year 2002 in the band combination RGB = NIR, SWIR, RED. Each column contains the nine images of each quarter of the year, in chronological order from top to bottom.

11.8 Case study: Flood modelling: Nam Chun (Thailand)

11.8.1 Introduction to case study

In this section we focus a problem that occurs all over the world: flooding. On every continent, floods cause damage and kill people, so one can quite safely say that floods are the most recurring, widespread and disastrous of all natural hazards. The occurrence of most floods is, of course, highly correlated with the meteorological conditions in an area, and since these are difficult to forecast with a high degree of accuracy, it is also difficult to predict flood events. But even if meteorological conditions—especially precipitation—are known, it is no trivial task to forecast floods, let alone to predict their severity. To be able to so, additional information is required, not only regarding the precipitation but also with respect to the area that is likely to be affected. For instance, with precipitation one needs to know the form in which it will fall (rain or snow, . . .), the total amount that will fall, its intensity and where exactly it will fall. In short, the precipitation must be spatially-dynamically characterized, where spatial stands for the geographic domain (where) and dynamic for the development over time (when). But also the area that receives the precipitation must be known: what is its size, how steep is it, what is the shape of the watershed that feeds its river system, what are its vegetation and soil characteristics, and how much water is already present to start with? In order to be able to predict how much water will be at a certain place at a certain time, we must enter the world of spatial-dynamic modelling.

Spatial-dynamic modelling

Most GISs have limited dynamic modelling capabilities, especially when it comes to problems where the output of a given time-step becomes the input of the next time-step. In most cases, therefore, a GIS is used in parallel with, i.e. in addition to, a dedicated spatial-dynamic model; the GIS is used to prepare the spatial input data that the model requires and to analyse and further elaborate the model's spatial output (see Figure 11.19). Of course there are exceptions to this, such as the dynamic GIS PCRaster developed by the University of Utrecht and some advanced models that contain basic GIS functionalities. In our case study, we will use, in sequence, two spatial-dynamic models parallel to a GIS. The first model is a rainfall-runoff model that is used to predict discharge at the outlet of a catchment—the upstream part—as a function of the rainfall and the characteristics of the catchment. The discharge predictions become the input for the second model: a 1D/2D hydraulic model that simulates the flow of water over (and through) complex topography in the downstream part of a catchment. It is used to assess the spread of flood water in order to estimate the consequences of flooding in the affected parts downstream.
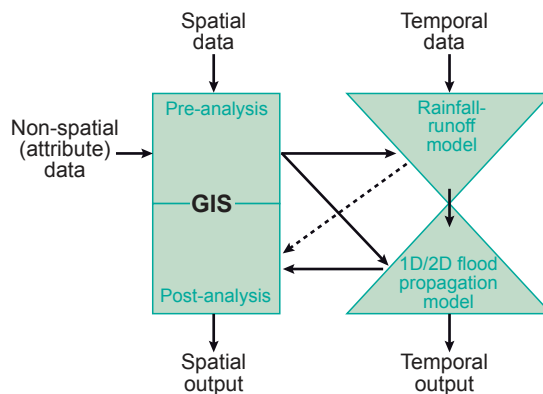
Figure 11.19
Parallel use of GISs and
spatial-dynamic models

The Nam Chun study

On 11 August 2001 the typhoon Usagi passed over central Thailand bringing with it intense and prolonged rainfall. This resulted in numerous flash floods and landslides. Approximately 120 people died in this event and over a 1000 people were made homeless. Very quickly people blamed the extensive and uncontrolled deforestation of central Thailand as the main cause of the widespread destruction the typhoon caused. The province of Petchabun in central Thailand—see Figure 11.20—was one of the worst affected areas. The central event in this disaster was the occurrence of a flash flood that originated in the Nam Chun watershed and caused extensive flood damage downstream on the Pa Sak flood plain.

The study described here was carried out to check to what degree had deforestation affected the generation of the flash flood that occurred and to establish to what extent would reforestation be a useful and effective measure for preventing a repetition of such disastrous flooding.

Figure 11.21 shows the main spatial characteristics of the watershed: an upstream, mountainous part (left-hand side), with relatively steep slopes and deeply incised river valleys; and a downstream part (right-hand side) with relatively flat topography, exhibiting a gentle gradient of approximately 1–2% towards the Pa Sak River, to which the Nam Chun River contributes. The area of the upstream and downstream parts of the watershed is in total approximately 92 km^2.

The upper part of the watershed consists of two parallel sub-catchments of a general west-northwest to east-southeast orientation that have their confluence near where the Nam Chun River enters the Pa Sak valley proper. In terms of lithology, the upstream part consists of uplifted Triassic sedimentary rocks, mainly conglomerates, sandstones and shales. The downstream part consists mainly of Quaternary colluvial and alluvial deposits. The vegetation in the upstream part can be characterized by degraded and disturbed forests on the steeper and higher slopes. In the lower parts, farmers have encroached upon the forest to cultivate maize and other food crops such as beans, cabbage and tamarind. A significant part of the upper catchment is covered by fallow grasslands.

Of the downstream part, almost the entire area is used for agriculture: farmers grow rice in the rainy season (May–September) and tobacco, cucumber and maize in the dry season (October–April). Tree crops such as coconuts, mangos and tamarind are also grown there. The average annual rainfall is 1066 mm per year from an annual average of 120 rainy days, which are concentrated in the period May to September. The average maximum temperature is 34 °C (ranging between 37 °C in April and 31 °C in December.

11.8.2 Surface-runoff modelling in the upstream part

In the study, a distributed erosion model was applied to quantify the amount of runoff in the upper catchment and to obtain hydrographs at the outlet into the Pa Sak valley. This model, the Limburg Soil Erosion Model (LISEM), was used because it takes into account the effect of land cover and soil characteristics in a spatial way. This meant that all model parameters could be represented as maps (in raster format), which not only allowed us to assess the consequences of, for example, land cover changes on the shape of the hydrograph, but also to identify where land cover changes would have the most significant effect. This last capability is important because it offers spatial planners a tool for prioritizing areas where mitigation efforts are likely to have the greatest effect. The reader should note that LISEM is a combined hydrological and erosion model, even though the erosion part was not used in this study: LISEM was applied only as a distributed hydrological model.

Figure 11.20
Location of the Nam Chun
watershed in North-Central
Thailand.

LISEM is a physically-based hydrological and soil erosion model for simulating hydrology and sediment transport during and immediately after a single rainfall event, and it may be applied in catchments of sizes ranging from 1 ha up to approximately 100 km². The model, developed by Utrecht University's, Department of Physical Geography, simulates the effects of both current land use and soil conservation measures on surface runoff and sediment transport. LISEM comprises two basic processes: a water part and an erosion part (Figure 11.22). The hydrological processes incorporated in the model are rainfall, interception, surface storage in micro-depressions, infiltration, vertical movement of water in soil, flow over land and channel flow. The model also incorporates the influence on hydrological and soil erosion processes of compaction (including the influence of tractor tracks), small paved roads, field strips and grassed waterways.
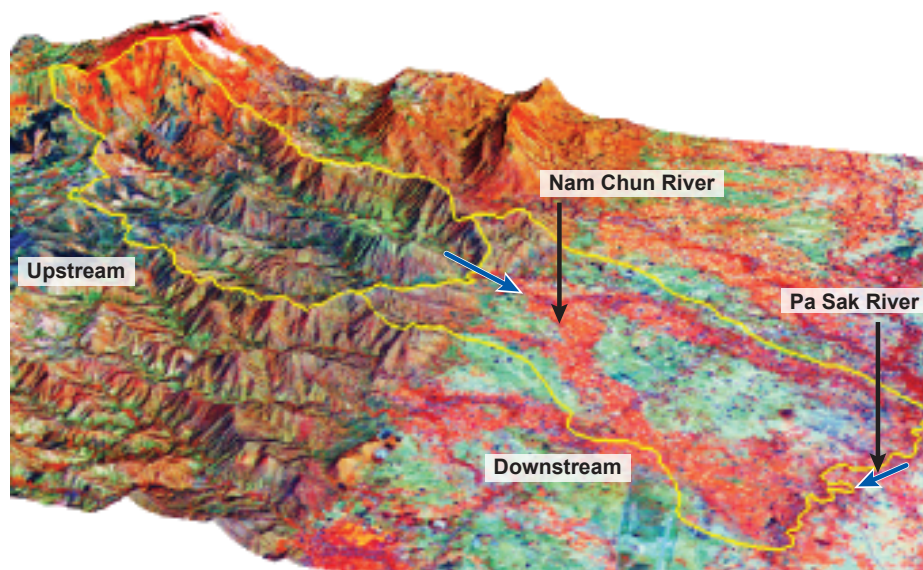
Figure 11.21
A 3D representation of an ASTER false colour composite of the Nam Chun watershed; scale varies in this representation. The view is towards the North.

For the runoff process, rainfall is the basic input. Interception by crops and vegetation is simulated by regarding them as a storage compartment that is to be subtracted from the rainfall. The remaining rainfall then reaches the soil surface, where it can infiltrate or be added to the surface storage. Since LISEM is a storm-based model, the infiltrated water is considered as a loss in the sense that it cannot resurface.

Infiltration can be simulated using one of several available equations, among them those developed by Green & Ampt, Holtan and Richards. In our study we used the Green & Ampt model. Surface storage can be considered as a mini reservoir in which water is stored (think of small ponds and puddles) until a threshold is exceeded. Then overland flow will occur. The flow velocity can be calculated using a combination of Manning's formula and the kinematic wave equation. The flow is directed over the terrain along the local drainage direction, which can be derived from the relevant DTM.

Since LISEM is a process-based model it required a significant amount of input data. All the required input data were derived from three base maps: 1) the digital elevation mode; 2) a soil unit map; and 3) a land cover map. The digital elevation model was derived from the topographic map (Land Development Department, Thailand, based on aerial photographs 1:25,000). The soil unit map was derived from a study by Solomon [107], and values for the infiltration parameters (Ksat) were obtained by Prachansri [93]. A digital land cover map of the study area was produced in 2004 by the Land Development Department of Thailand. The accuracy of the map was found by Prachansri [93] to be 72%, which we considered to be acceptable for the purpose of our study. As LISEM is grid-based, all input maps used in the model were converted into a raster format having a grid size of 30 m × 30 m. Table 11.2 shows a number of the parameters (collected by Prachansri [93]) required for running the model.
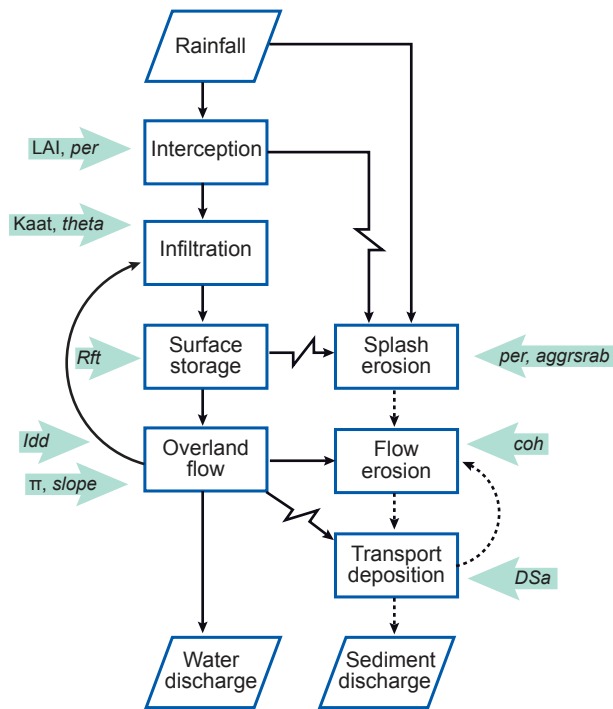
Figure 11.22
Simplified flow chart of the
LISEM model.
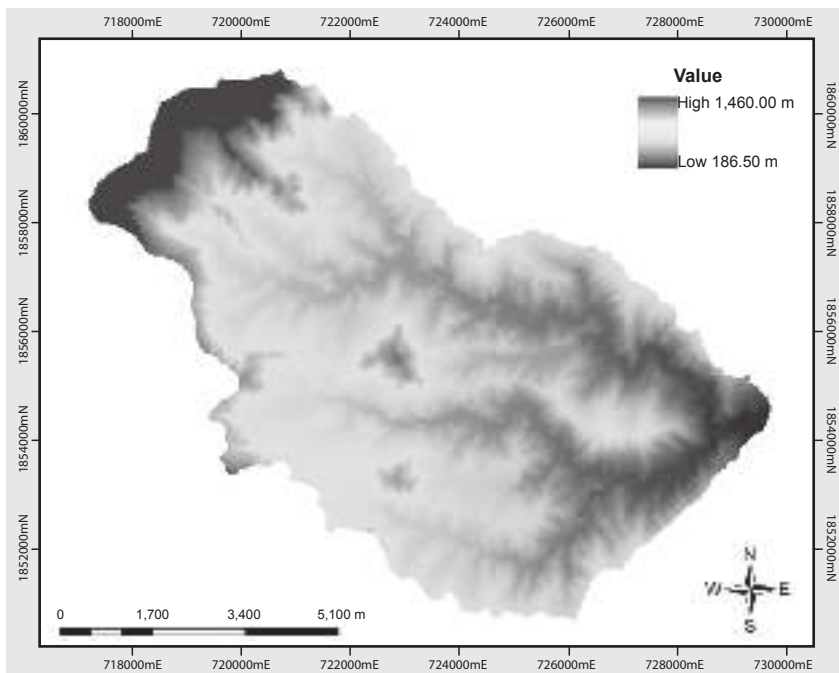
Figure 11.23
Digital Elevation Model
(DEM) used in the LISEM
model

Table 11.2
Input data for LISEM
version 2.39, with the use of
the Green & Ampt infiltration
model.

Parameter	Name	Method	Unit
Catchment characteristic			
Local drain direction	LDD.map	derived from DEM	-
Catchment boundaries	AREA.map	derived from DEM	-
Area covered by rain gauges	ID.map	field observation	-
Slope gradient (sine of slope angle)	GRAD.map	derived from DEM	-
Location of outlet and sub-outlets	OUTLET.map	derived from DEM	-
Rainfall data	ASCII	field observation	mm/h
Interception			
Fraction of soil covered by vegetation	PER.map	field observation	-
Leaf area index	LAI.map	derived from PER.map	-
Vegetation height	CH.map	field observation	m
Infiltration			
(Green & Ampt)			
Saturated hydraulic conductivity	KSAT1.map	field measurement	mm/h
Saturated volumetric soil moisture content	THETAS1.map	field measurement	-
Initial volumetric soil moisture content	THETAI1.map	field measurement	-
Soil water tension at the wetting front	PSI1.map	derived from literature	cm
Soil depth	SOILDEP1.map	field observation	mm
Surface storage			
Random roughness	RR.map	derived from literature	cm
Width of impermeable roads	ROADWIDT.map	field observation	m
Overland flow			
Manning's roughness coefficient	N.map	derived from literature	-
Local drain direction of channel network	LDDCHAN.map	derived from ldd.map	-
Channel flow			
Local drain direction of channel network	LDDCHAN.map	derived from ldd.map	-
Channel gradient	CHANGRAD.map	derived from grad.map	-
Manning's *n* for the channel	CHANMAN.map	derived from literature	-
Width of channel scalar	CHANWIDT.map	derived from ldd.map	m
Channel cross-sectional shape	CHANSIDE.map	field observation	-

Model calibration and validation

Many authors have demonstrated the need to calibrate process-based models to achieve acceptable levels of predictive quality. In the case of hydrological models, typically these are calibrated using data measured at the outlet of the relevant catchment. Differences between observations and simulated modelling results can be basically attributed to four different sources of error:

1. errors in the meteorological input data;

2. errors in the recorded hydrological observations;

3. errors and simplifications inherent in the model's structure; and

4. errors resulting from the use of non-optimal parameter values.

During the calibration step, only errors from non-optimal parameter values can be addressed.

In our case study, the results of the LISEM model were evaluated using six selected rainstorm events (Tables 11.7–11.8). Three of these served as a calibration set, to optimize the parameter settings. The other three rainstorms served as a validation set, to test the predictive qualities of the optimized model.

The model results were calibrated primarily on peak discharge, but also on the general shape of the hydrograph. The simulated hydrograph was visually compared with the measured data and the two parameters were used to calibrate on peak discharge:

1. Saturated hydraulic conductivity (Ksat), which determines infiltration rate and amount of runoff; and

2. Surface roughness coefficients. Table 11.3 shows the results for the peak discharge and Figure 11.24 the results for the general shape of the hydrograph. The parameter values obtained after the calibration step can be found in [93].

Events	Rainfall (mm)	Peak discharge ($m^3 s^{-1}$)	
		Observed	Simulated
Calibration			
60905	52.58	37.5	37.90
18095	18.43	13.00	13.40
260905	29.49	31.45	32.99
Validation			
70905	16.60	4.2	4.72
90905	10.61	1.02	1.30
120905	26.74	15.37	23.81

Table 11.3
Observed and simulated peak discharge in the Nam Chun catchment.

Runoff rate for different land cover types

The model enabled us to assess the contribution to runoff from sub-areas within the catchment. The highest volume of surface runoff was predicted for maize cultivations, with an average value of 482×10^{-3} m^3 s^{-1}; the lowest volume predicted was for forest areas, with an average of $13.3 \times 10^{-3} m^3 s^{-1}$. On the whole, agricultural areas, comprising cornfields, mixed crops and orchards, show approximately 16 times higher
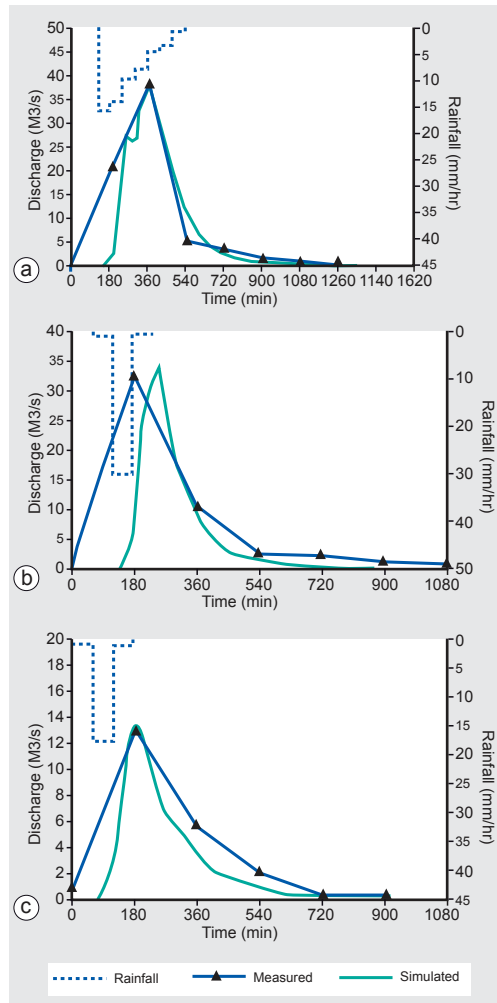
Figure 11.24
Measured and simulated
discharge in the Nam Chun
catchment on 6 September
2005 (top-left), 18 September
2005 (top-right) and 26
September 2005 (bottom).

rates of surface runoff than non-agricultural areas (comprising forest, degraded forest and grassland areas): $440 \times 10^{-3}\,\mathrm{m^3\,s^{-1}}$ vs. $27.82 \times 10^{-3}\,\mathrm{m^3\,s^{-1}}$ per pixel of $30\,\mathrm{m} \times 30\,\mathrm{m}$.

This enormous difference can be attributed to the combination of higher hydraulic conductivity, greater surface cover and higher surface roughness values in the non-agricultural areas, resulting in high infiltration rates, which in effect reduce surface runoff. The spatial and temporal distribution of surface runoff as predicted by the calibrated model is presented in Figure 11.25.

Time-step 60

Time-step 120

Time-step 180

Time-step 240

Time-step 300

Time-step 360

Time-step 420

Time-step 480

Increasing surface runoff (1s^{-1})
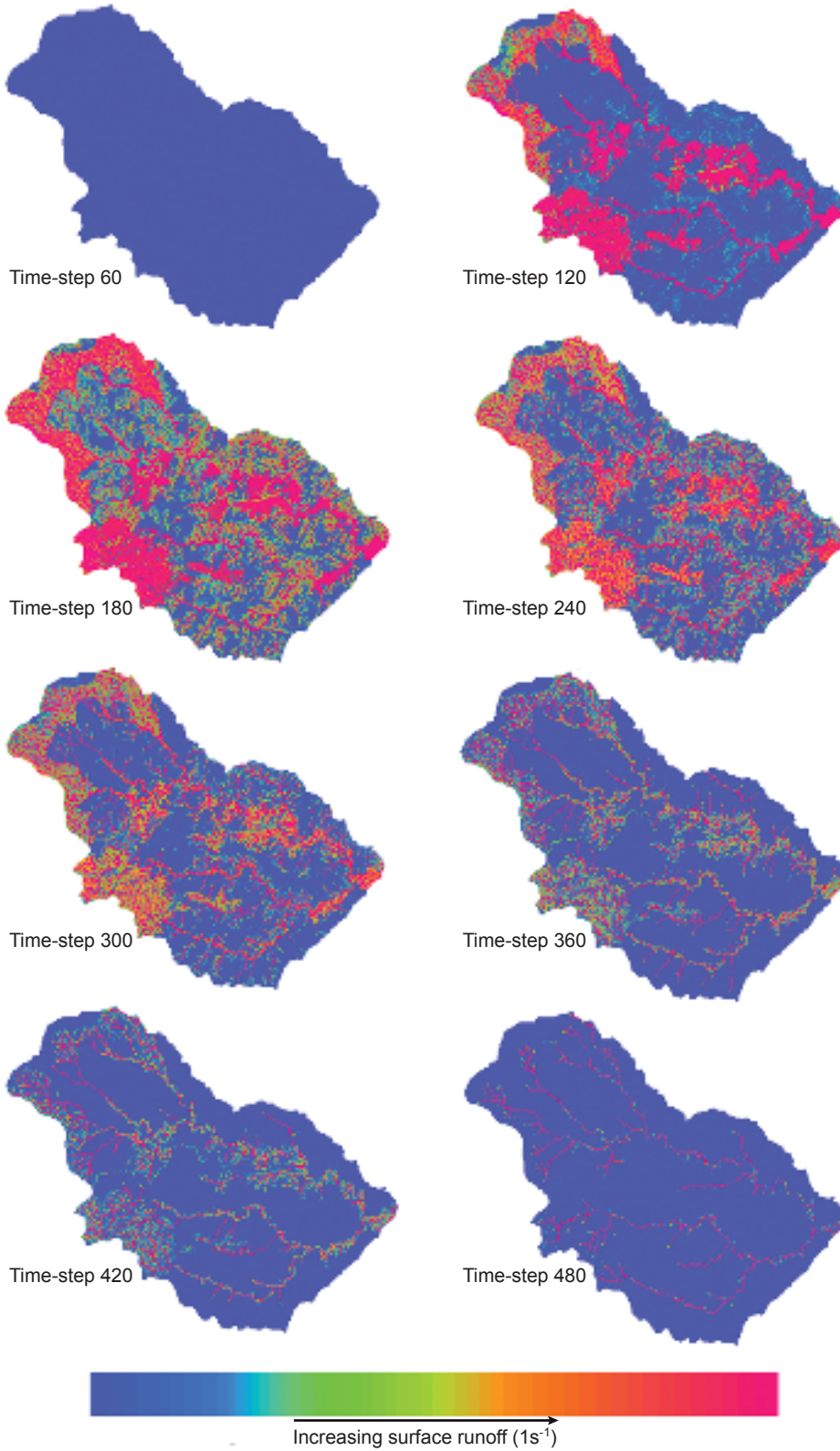
Figure 11.25
Spatial and temporal distribution of surface runoff (time-step = 1 minute).

Scenario generation

Our objective in the Nam Chun study was to evaluate the effects of different land use scenarios on the rate of predicted surface runoff in the catchment. Jetten et al. [49] state that although models may not be able to accurately predict future events, they can be used to compare different scenarios. In scenario studies, the same uncertainty about input data applies to all scenarios and one can, therefore, assume that the differences resulting in the different simulations are in fact a consequence of the scenario changes applied. Therefore, in order to evaluate the effects of different land use scenarios, three land cover scenarios were developed:

- Base scenario: actual situation, i.e. as per August 2001;

- Scenario A: change the entire catchment to forest;

- Scenario B: Change land use to corn cultivation before harvest; and

- Scenario C: Change land use to corn cultivation after harvest (i.e. bare soils).

The results of the scenario simulations are shown in Table 11.4.

Table 11.4
Summary of change in peak runoff and its time of occurrence for a selected event (No. 060905) in relation to different land use scenarios: (A) entire catchment forested; (B) change from forest to corn fields before harvest; (C) change from forest to corn fields after harvest. Change in peak discharge is given compared to that of the actual situation. Peak arrival time for the actual situation was 3.5 h.

Scenarios	Peak discharge(m^3s^{-1})	Change in peak discharge	Peak arrival time difference(h)
Actual situation	37.9		
A	9.3	-76%	+2.0
B	184	+385%	-2.0
C	194	+412%	-3.0

Under Scenarios B and C, total discharge was predicted to increase to approximately 400% compared to that under actual land use (base scenario). This indicates that cultivation of corn increases the amount of surface runoff and that expanding areas of agricultural activity in the watershed will result in a drastic increase in river discharge at the catchment's outlet. Figure 11.26 shows the predicted discharges for the three land use scenarios.

11.8.3 Flood propagation modelling in the downstream part

In previous subsections we have demonstrated how a hydrological model can be used to estimate runoff from a given catchment for various scenarios. In itself this is useful, but in practice it is not in the catchment itself where most flooding problems occur. Flooding is usually greatest on the relatively flat terrain beyond the outlet of the mountainous catchments, where we usually find higher concentrations of human population and property.

In this subsection we demonstrate how to assess the consequences of this runoff in the downstream parts. In order to quantify the effect of upper catchment runoff, a combined 1D/2D hydraulic model was used to model the propagation of the flow over flat terrain. To do this we chose to use the SOBEK modelling suite, which combines one-dimensional channel flow with two-dimensional overland flow.
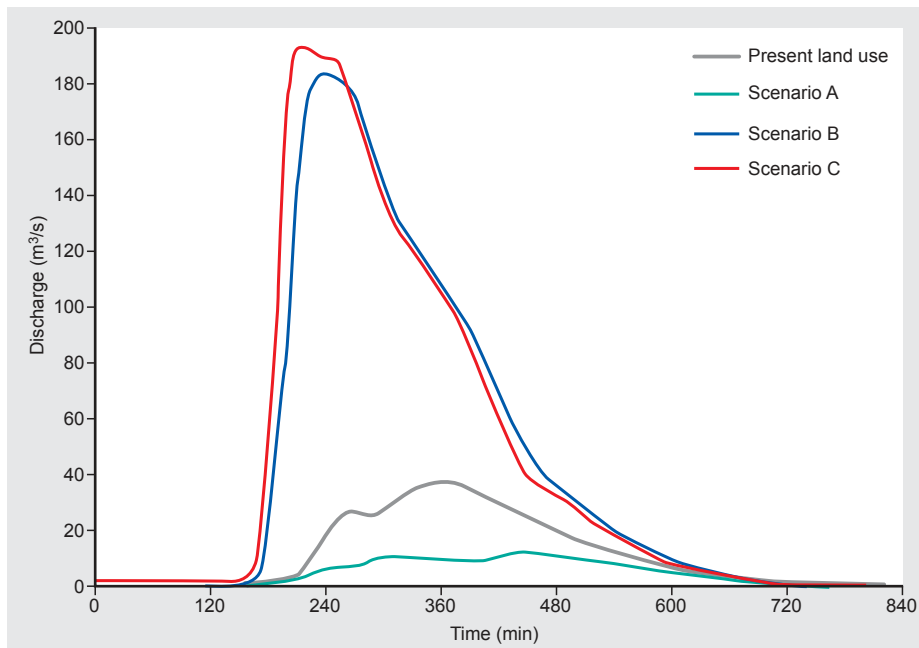
Figure 11.26
Predicted hydrograph for three land use scenarios.

Model data input

Four main types of input data were needed for the model: 1) The digital terrain model, to represent the natural and man-made topography (this includes flow-affecting structures such as embankments and road networks); 2) surface roughness data, to represent the resistance of different vegetation types along the river channel and the flood plain on the water flow; 3) river cross-sections, to represent the shape of the rivers; and 4) the boundary conditions, which include the incoming discharge at the upstream boundary and water levels, or a rating curve, at the downstream boundary. Note that in the model as it was used in this study, there was no direct rainfall in the downstream area: it was assumed that all surface water came from the upstream part of the catchment.

Surface topography

By far the most important input for the model are data that accurately represent the surface topography. Especially for the Nam Chun study, a detailed topographical survey was conducted by the mapping division of the Thai Land Development Department [93], producing 1 m contour lines and spot heights with an accuracy of 0.1 m (Figure 11.27). From this data, a DTM was derived with a grid size of 10×10 m. During a field survey, man-made objects that affect flow, such as embankments, were added to the DTM.
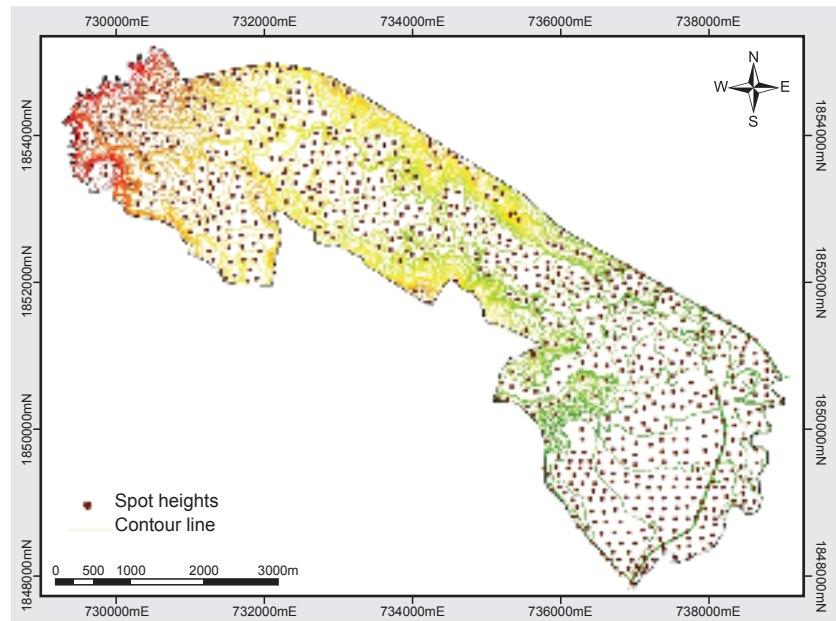
Figure 11.27
Spot heights and contour lines of the Nam Chun flood plain.
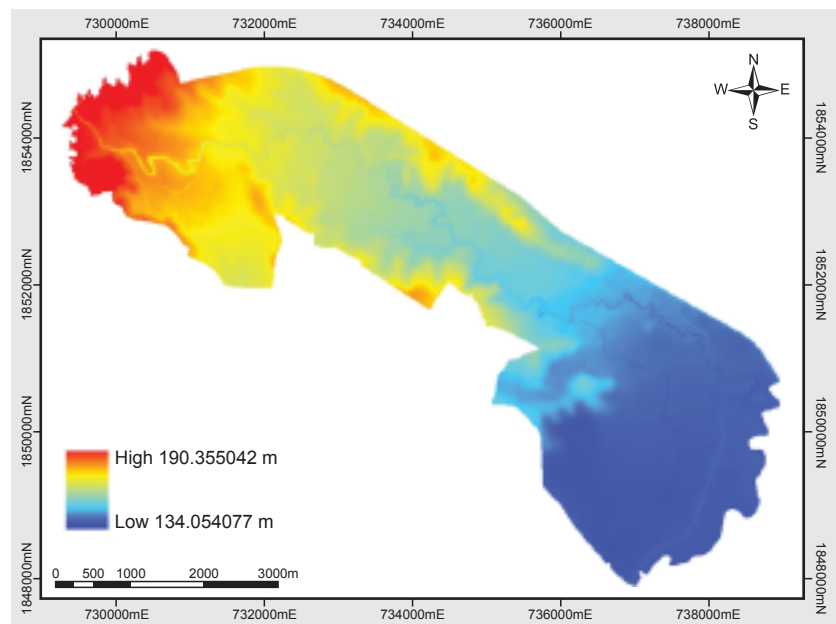
Figure 11.28
Digital Elevation Model of the Nam Chun flood plain.

Surface roughness map

In flood modelling, surface roughness represents the resistance that the water experiences as it flows over the surface of terrain. Surface roughness is strongly related to land cover: smooth, non-vegetated surfaces offer little resistance, whereas dense forests have high friction values. The parameter used for expressing surface roughness is Manning's roughness coefficient. It is hard to measure this coefficient under normal conditions, so usually tabulated values as a function of land cover are used. During calibration procedures, Manning's roughness coefficients are frequently adjusted to improve model performance.

The land cover map that was used for the downstream part of the Nam Chun study was derived using visual interpretation of aerial ortho-photos at a scale of 1:25,000, and the land cover was classified according to the classification standard of Land Development Department of Thailand. The land cover map was further improved during a field survey. The final map is shown in Figure 11.29. This map was transformed into surface roughness values using values of Manning's coefficients taken from the literature; see Table 11.5.
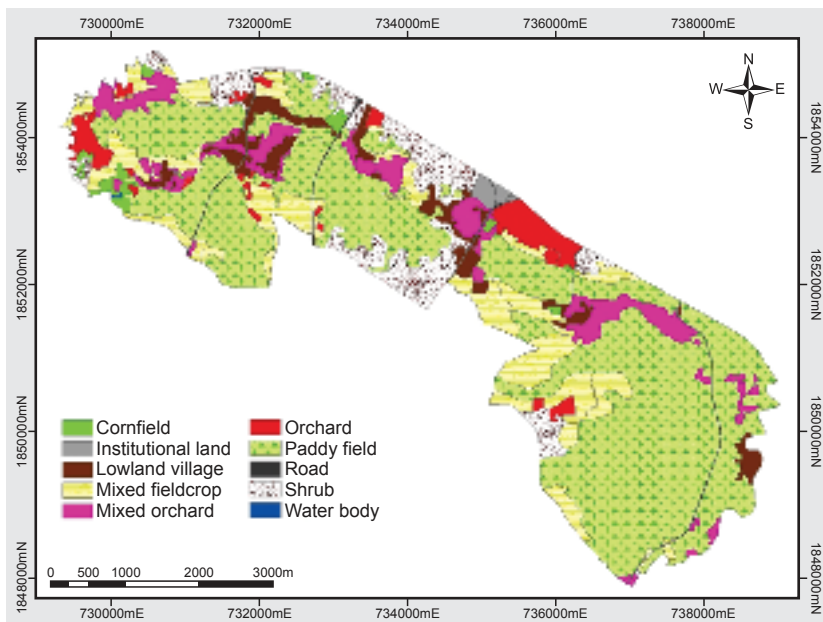
Figure 11.29
Land cover types on the Nam Chun flood plain.

River cross-sections

In this study, cross-section data were derived from the DEM and by surveying during the fieldwork. The channel width and slope of the river bank were measured and visually estimated. River-bed elevations were obtained directly from the DEM.

Boundary conditions

The upstream boundary condition is the discharge prediction from the catchment study as described in Subsection 11.8.2 and Figure 11.26. In our study, not only did we make discharge predictions for the three land use scenarios in the upstream catchment; we also made predictions for rainstorms of different return periods: 2, 10, 20 and 50 years. For example, a rainstorm with a return period of 50 years is an event that, on average, occurs once every 50 years. The return period is a measure of the severity of

Land cover types	Manning's coefficient
Cornfield	0.045
Shrub	0.040
Mixed field crop	0.035
Mixed orchard	0.150
Orchard	0.100
Paddy field	0.100
Institutional	0.001
Lowland villages	0.150
Roads	0.001
Water body	0.033

the storm: a longer return period signifies a more severe storm. Hydrographs for all three scenarios are shown in Figure 11.30 a,b. The downstream boundary consists of a time series of water levels observed in the Pa Sak river.

Model output

SOBEK generates several output files, such as maps at predefined time-steps that contain information on water depth and flow velocity; time series of water depth, discharge and flow velocity at predefined locations; and an animation file that shows the progression of the flood. In the Nam Chun study, we used maximum water depths and maximum velocities for the flood-hazard mapping.

Calibration of the model results

In this study, the model was calibrated by comparing water-depth maps with the water depths that were obtained from interviews during fieldwork in 2001, after typhoon Usagi. Unfortunately, there was no information available on the extent of flooding. During the fieldwork, 50 flood-depth points were collected through interviews with the local population. The parameter that was used to minimize differences between the modelling results and observed flood data was the surface roughness coefficient. After a limited number of trials, a set of optimum values was obtained that were used during further analysis. For more information regarding the calibration step, please check [93].

Modelling results for the three land use scenarios

The results shown in Table 11.6 indicate that for a rainstorm with a return period of 2 years under actual land cover the average inundation depth in the upstream part is 0.35 m, although in some places a maximum depth of 0.89 m would be reached. Only 8% of the territory would be flooded in the downstream part (Figure 11.31). Under scenarios in which the upstream catchment is completely transformed into corn fields (Scenarios B and C), the maximum flood-water depth downstream doubles and the spatial extent of flooding on the flood plain more than triples. If the entire upland area is covered by forest (Scenario A), there would be no flooding at all downstream.

Total area 25 km^2	Actual situation	Scenario A	Scenario B	Scenario C
Flooded area, km^2	1.92	0.07	7.27	7.65
Total flood volume, 10^6m^3	0.29	0.00	2.69	3.20
Average depth, m	0.35	0.14	0.96	0.98
Maximum depth, m	0.89	0.41	2.21	2.25
Maximum velocity, ms^{-1}	1.11	0.12	5.51	5.8
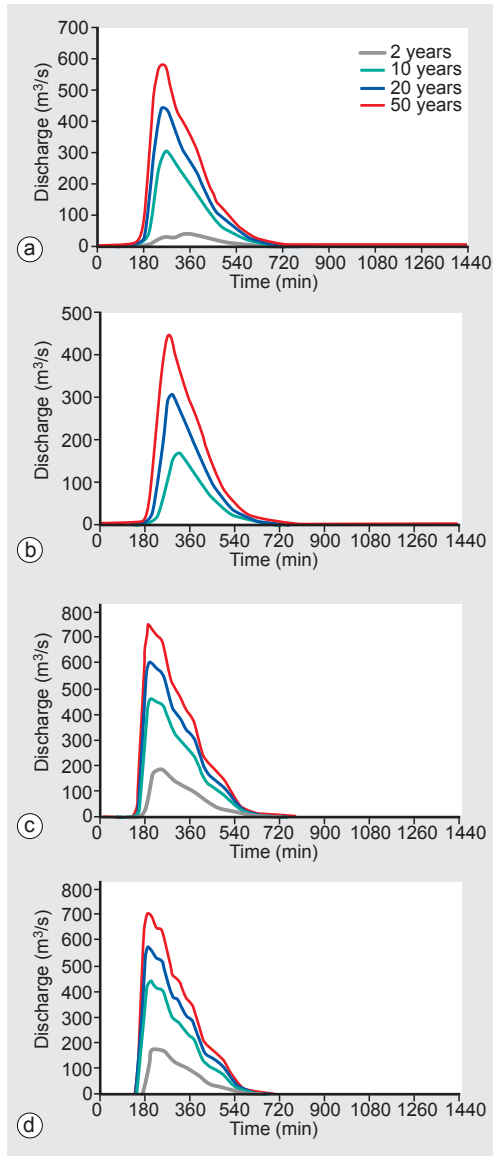
Figure 11.30
Boundary condition for Nam Chun upstream for 4 return periods: (a) actual land use, (b) Scenario A, (c) Scenario B and (d) Scenario C.

11.8.4 Example: flooding following a rainstorm with a 20 year return period

For a rainstorm with a 20 years return period, the results of the model simulation (Table 11.7) indicate that under the actual situation 55% of the area would be inundated. For the Scenarios A, B and C, the inundated area would be 43%, 64%, 63%, respectively. Under Scenario B, the total volume of flood water would be $19.59 \times 10^6\,\mathrm{m}^3$, which is twice as high as under the actual situation. For Scenario C, the total of volume of flood water was lower than for Scenario B. This can be explained by the distribution of discharge over time. Under scenario C, for a rainstorm with a 20 year return period (annual probability of occurrence 5%) the discharge is higher at the beginning but after two hours it subsides to levels below that of Scenario B, thus affecting the total volume of the flood water under Scenario B.

Table 11.7
Summary of flooding characteristics under the three land use scenarios for a rainstorm with a 20 year return period.

Total area 25 km²	Actual situation	Scenario A	Scenario B	Scenario C
Flooded area, km²	13.87	10.77	16.00	15.77
Total flood volume, 10^6m³	9.47	5.21	19.59	14.66
Average depth, m	1.53	1.24	1.87	1.79
Maximum depth, m	3.40	2.96	3.90	3.97
Maximum velocity, ms^{-1}	6.04	5.77	6.99	6.92

Maximum water depth

Figure 11.31 and Table 11.8 show that the maximum water depth maps for scenarios B and C are quite similar, with only a small area at the lower part of the map for Scenario B having a greater maximum water depth than for Scenario C. In contrast, the combined maximum water depth class 0.2–1.0 m and flood extent under Scenario C is greater and larger than under Scenario B. These depths are found near the apex of the Nam Chun because under Scenario C the water propagates and inundates in the upper part of the river—then there is less water to drain to the river's lower portion. In Scenario A, the maximum depth of most of the floodwater does not reach 2 m, yet under Scenarios B and C areas of 18% and 9%, respectively, would be flooded by water with depths greater than 2 m; maximum water depths greater than 3 m cover 6% of the surface area under Scenario B. Under Scenario A, 88% of the flooded area has a water depth of less than 1 m.

Table 11.8
Surface area flooded (% of total area) per water depth class for a rainstorm with a 20 year return period.

Water depth, m	actual situation	Scenario A	Scenario B	Scenario C
0 - 0.2	19	28	10	11
0.2 - 0.5	28	33	14	19
0.5 - 1.0	29	27	31	34
1.0 - 2.0	22	11	27	27
2.0 - 3.0	2	1	12	8
> 3.0	0	0	6	1
Total (km²)	100 (13.87)	100 (10.77)	100 (16.00)	100 (15.77)
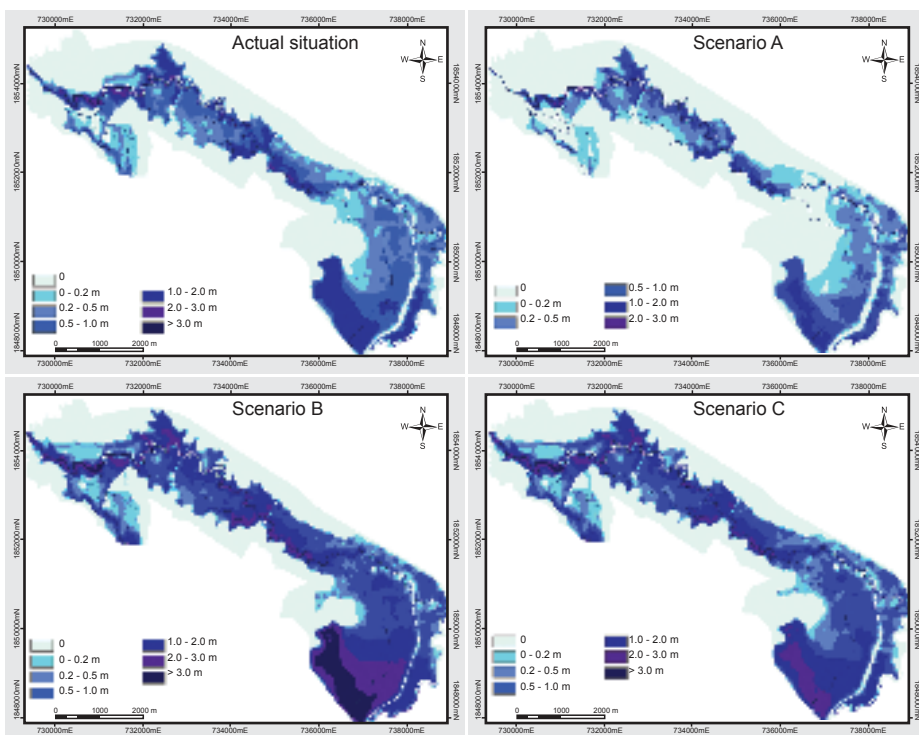
Figure 11.31
The spatial distribution of maximum water depth for the three land use scenarios and actual land use for a rainstorm with a 20 year return period.

Maximum flow velocity

Figure 11.32 and Table 11.9 show the distribution of the maximum flow velocity for the present situation and scenarios A, B and C. For most of the inundated area the maximum flow velocity does not rise above 0.50 m s^{-1}. In scenarios B and C, the maximum velocities of water flow that are less than 0.50 m s^{-1} occur on 97% and 96%, respectively, of flooded area; 4% of the area has a flow velocity greater than 0.50 m s^{-1}. The highest maximum water flow velocities are to be found in the top part of the downstream area (near the apex). In scenario A, for most of inundated area the flow velocity of flood water is lower than 30 cm s^{-1}. This means that if the upland catchment were completely forested, it would reduce the maximum water flow velocity in the downstream area.
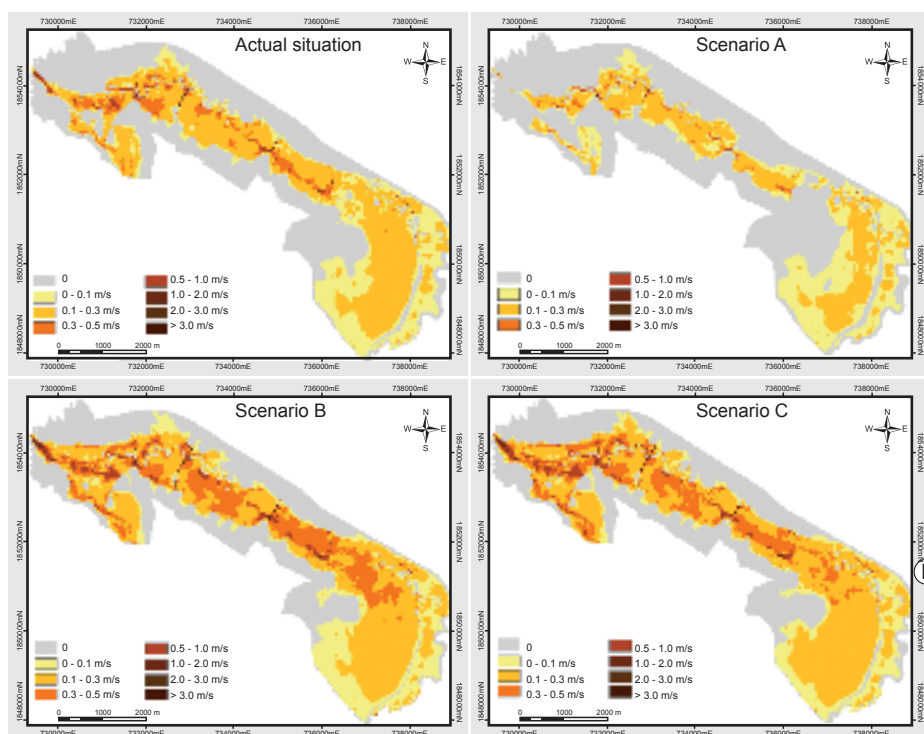
Figure 11.32
The spatial distribution of maximum flow velocity for the three land use scenarios and actual land use for a rainstorm with a 20 year return period.

Table 11.9
Surface area flooded (% of total area) per water velocity class for a rainstorm with a 20 year return period.

Water velocity, ms^{-1}	Actual situation	Scenario A	Scenario B	Scenario C
0 - 0.1	31	50	25	20
0.1 - 0.3	59	47	51	59
0.3 - 0.5	8	3	21	17
0.5 - 1.0	2	0	3	4
1.0 - 2.0	0	0	0	0
2.0 - 3.0	0	0	0	0
> 3.0	0	0	0	0
Total (km^2)	100 (13.87)	100 (10.77)	100 (16.00)	100 (15.77)

11.8.5 Flood-hazard mapping

Flood hazard is the probability of occurrence of a potentially damaging flood event of a certain magnitude within a given time period and area. As part of our case study, we created flood-hazard zone maps from the results of flood modelling simulations for rainstorms with return periods of 2, 10, 20 and 50 years. A different degree of hazard was assigned to each flooding frequency. Five categories of flood hazard were established for each scenario:

- areas with high flood hazard—high frequency of flooding, i.e. a return period of 2 years;

- areas with medium flood hazard—medium frequency of flooding, i.e. a return period of 10 years;

- areas with low flood hazard—low frequency of flooding, i.e. a return period of 20 years;

- areas with very low flood hazard—very low frequency of flooding, i.e. a return period of 50 years;

- areas with no flood hazard.

Figure 11.33 shows the flood hazard zone for each scenario and for the actual situation. A high level of flood hazard that applies to 31% of the area occurs when the whole upstream catchment area is turned over to corn fields (Scenarios B and C). In contrast, the area of high flood hazard becomes smaller if the upstream area is converted to forest. The area of medium levels of flood hazard (for rainstorms of 10 year return period) covers 36% of the terrain in the actual situation and under Scenarios B and C. Under scenario A only 28% of the area has a medium level of flood hazard. The areas of low flood hazard (for rainstorms of 20 year return period) cover only 3, 3, 2 and 1 km^2 under the actual situation, Scenario A, Scenario B and Scenario C, respectively. Thus we can conclude that flood hazard is reduced if the land cover upstream is entirely forest. By contrast, when the upland area is devoted to agriculture, the risk of flooding increases, as does the area of flood hazard.
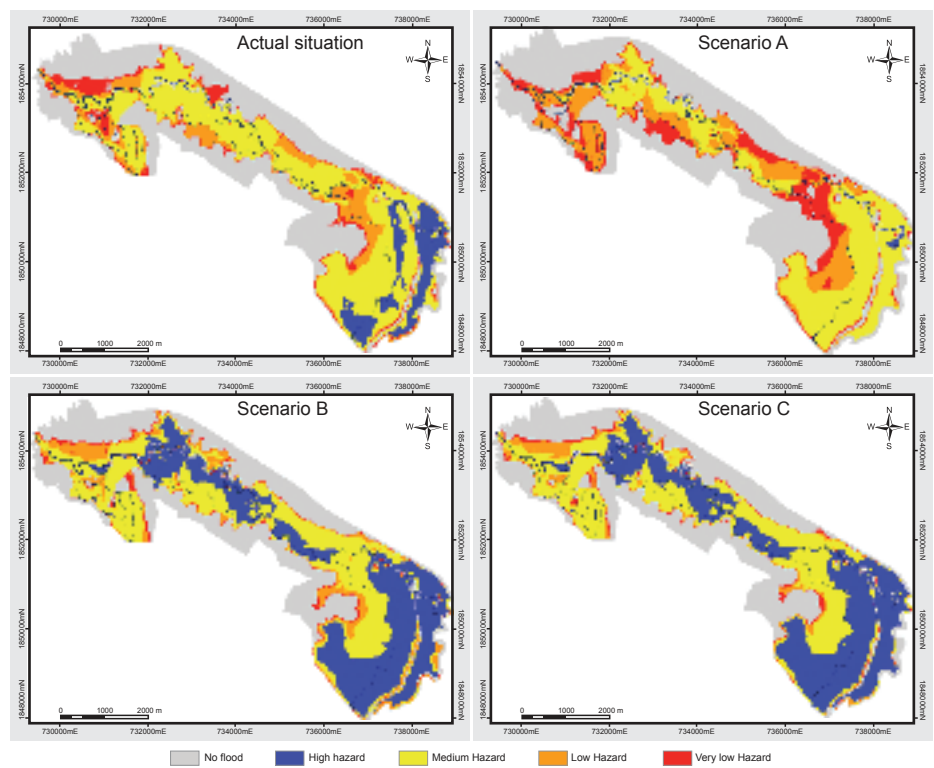
Figure 11.33
Flood-hazard mapping for the three scenarios of land use and actual land use.

11.9 Case study: Environmental management plan for the Lake Uromiyeh ecosystem, Iran

This case study illustrates how data integration at different levels was used in a complex project to support decisions about water allocation in a semi-arid region. The discussion here focuses on data integration issues without going into much detail about the project itself.

11.9.1 Project set up

Lake Uromiyeh (Figure 11.34) lies in the western part of Iran. Its basin covers about 54,000 km^2 and is made up of mountains and river flood plains. The climate is semi-arid, with an average annual rainfall of about 350 mm; low-lying areas (1200–1400 m amsl) receive about 250 mm per year and mountainous areas (around 3000 m amsl) more than 1000 mm, mostly as snow. Irrigated agriculture is the main economic activity in low-lying areas, exploiting both surface water and groundwater resources, while rain-fed agriculture is practised in the foothills of the mountainous areas.
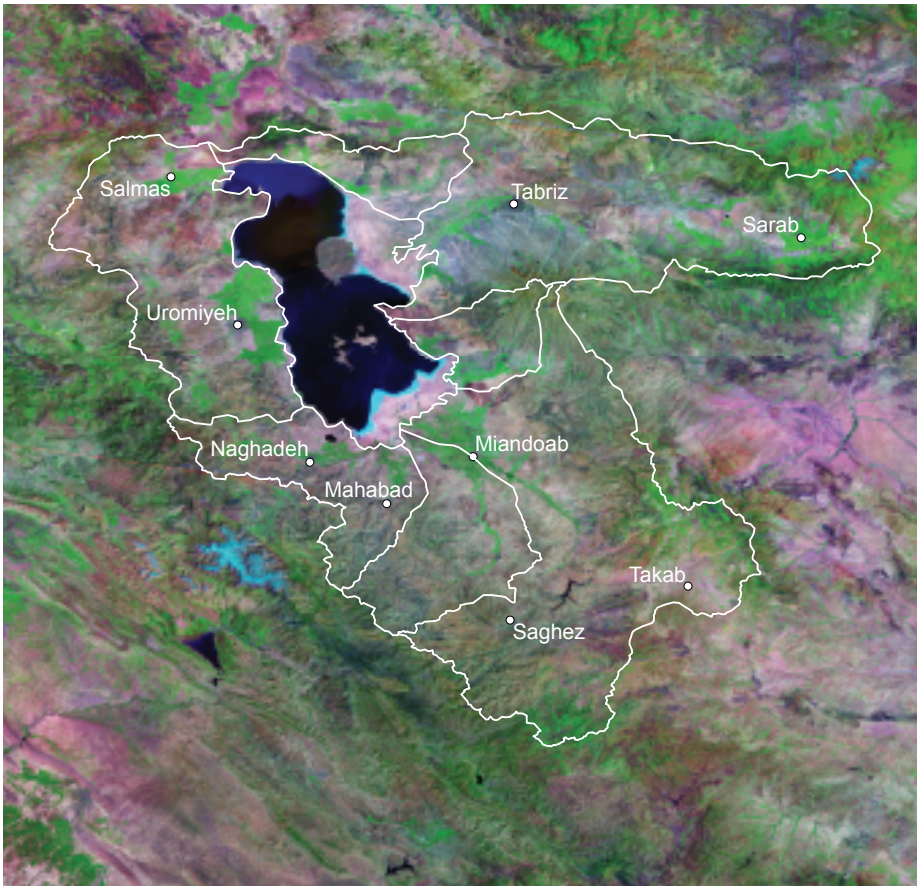
Figure 11.34
Lake Uromiyeh and its watershed; weather stations are sited at locations shown in the basin (source: U.S. Geological Survey Department of the Interior/USGS).

Lake Uromiyeh is shallow (6–8 m deep) and has no outflow, so any precipitation falling in the basin only leaves it through evaporation, either off the land or from the lake itself. The evaporating water leaves behind its dissolved salts, which at the time of the project had resulted in hypersalinity in the lake; the actual salinity depends on the amount of water in the lake, i.e. the lake's water level. These circumstances

have led to the development of a relatively simple ecology with salt-loving organisms populating every level of the food chain.

Falling water levels in the lake caused the shoreline to retreat, leaving behind deserts of salt and resulting in the complete collapse of much of the lake's flora and fauna. These events presented the authorities with a warning that better coordination of water use— involving all stakeholders in the basin—was needed.

The objective of the project was to develop an environmental management plan for the Lake Uromiyeh Basin in the framework of a cooperation between The Netherlands and Iran.

11.9.2 Decision-support system

As the success of the project required the cooperation of all sectors of the economy, as well as all groups of local inhabitants, the concept of integrated water resources management (IWRM) was applied. This entailed the involvement of stakeholders from agriculture, water management, industry, municipalities and water managers, together representing the governmental, private and non-governmental sectors.

The key tool in the integrated management was a decision-support system (DSS) that combined data from all related fields. IWRM, however, involves more than just running a DSS and making decisions on the basis of the outcomes it generates. Rather, it is the process by which the activities of all stakeholders are coordinated. The advantage of using a DSS is that it provides the possibility of testing outcomes of different decision schemes, i.e. of analysing different scenarios.

The DSS for Lake Uromiyeh incorporated a number of software tools, as shown in Figure 11.35.
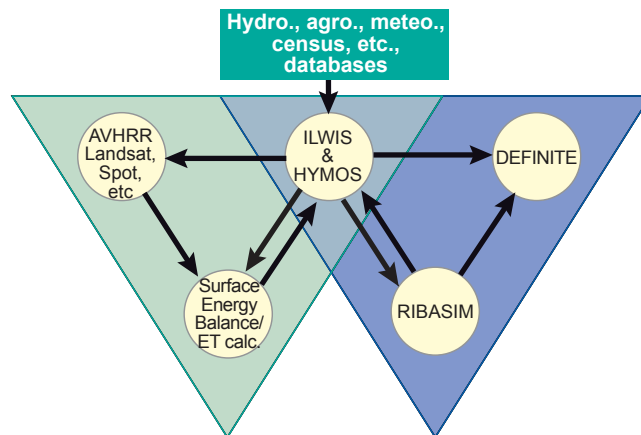
Figure 11.35
Logical structure of the Lake Uromiyeh DSS, indicating the major software tools and data types used.

The DSS comprised four major logical units:

1. external databases (green area in Figure 11.35), which did not form an integral part of the decision-support system;

2. a central spatio-temporal database (purple area) (a loose amalgamation of a GIS and hydrological data management software);

3. EO data and observation models (the overlap area between green and purple areas), which were used independently from the rest of the DSS. (In fact, this part processes input data for the decision support part of the system);

4. process models that exchanged data but were not fully integrated with the central database.

The software tools were not integrated in one computer system, but were used independently. A well-defined data flow scheme represented the logical framework of the DSS.

11.9.3 External databases

Several institutions from different Iranian ministries and departments provided data for the project. Hydrological, agronomic, meteorological, statistical, ecological and topographic data were merged in the central database. In many cases the data were not just copied from the external databases to the DSS: some conversions were made to meet the import requirements of the data storage systems.

11.9.4 Central spatio-temporal database

The central database used was not fully integrated to store the data in a homogeneous manner because the development of such an environment would have exceeded project funding. Instead, efforts were concentrated on storing the data in a practical structure: data for the process models were integrated into a special hydrological data management environment (HYMOS), while the data needed for the EO observation models were stored in an analysis-oriented GIS (ILWIS). This structure is illustrated in Figure 11.36; the special data integration tools and methods are indicated for the relevant steps.
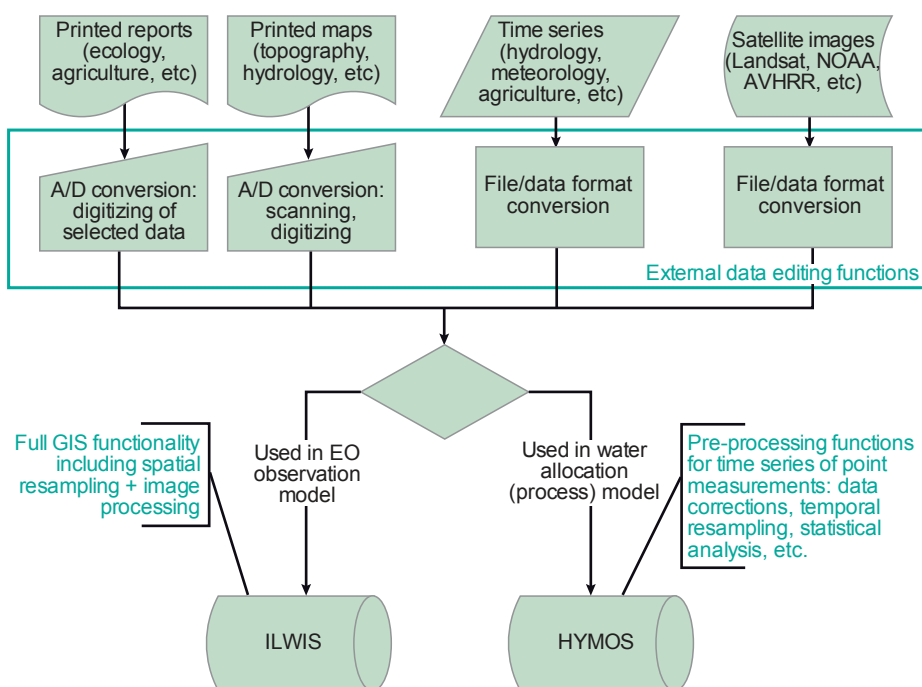
Figure 11.36
Data flows to the central spatio-temporal database of the DSS

An important characteristic of the main units of the central database was the wide range of data processing functions available. ILWIS was used mainly for the spatial operations[1], of which the most frequently used were:

---

[1]Not all the steps described here were originally carried out during the development phase of the central

- Re-projection of maps. For example, precipitation maps (Figure 11.37) were available originally in geographic coordinates, which were re-projected to the UTM mapping system.

- Spatial re-sampling for adjusting spatial resolution.

  – Re-sampling of coarse resolution to smaller grid size, i.e. densification. Precipitation maps are good examples of this here: the original maps of a course spatial resolution (about 10 km) needed to be densified, since several calculations used a grid of 1 km size. The densification method applied used an interpolation of regularly spaced data.

  – Re-sampling of finer resolution to coarser grid size, i.e. aggregation. This was used for generalizing the land cover/land use map of 28 m resolution that was provided by the Ministry of Agriculture and Jihad of Iran (originally classified from Landsat TM images) to the grid size of 1 km size. To do this a majority filter was used, which assigns the category of highest occurrence within the area of the aggregated cell to the whole cell.

- Spatial interpolation of irregularly-spaced point data; both simple and complex interpolators were used. A typical application was the interpolation of meteorological data (e.g. hours of sunshine) measured at six stations around the basin. For the interpolation, a complex geostatistical method was used that took into consideration both the spatial distribution of the variables shown in the IWMI Water and Climate Atlas of the region and the values actually measured at the stations.
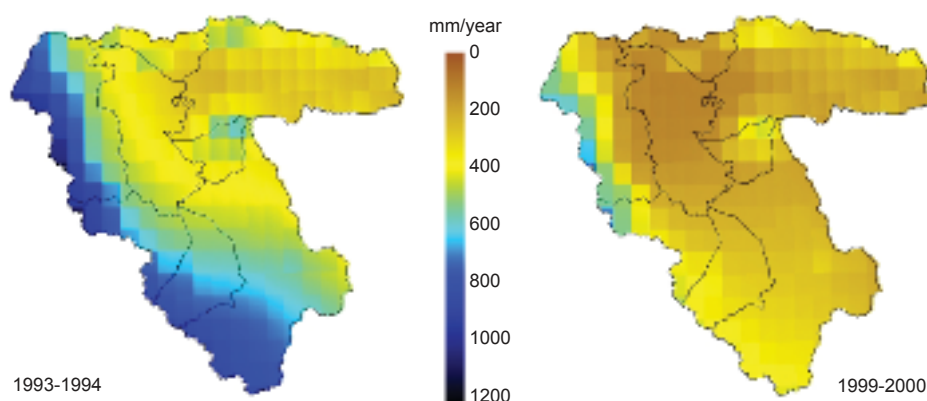
Figure 11.37
Precipitation maps of a wet hydrological year (1993–1994) and a dry hydrological year (1999–2000).

11.9.5 EO data and observation models

Earth observation was an integral part of the data analysis. Two major types of analysis based on Earth observation data were used:

- Satellite images were used for mapping the land cover and defining the major water users from this map (e.g. irrigated and rain-fed agriculture). This is the data flow related to the "Classification" observation model in Figure 11.38. Classification uses statistical methods for creating a link between the observed signal (the image) and the surface properties; the intermediate results were stored in the central spatio-temporal database (not shown in the figure).

spatio-temporal database. In its final form, however, the central database did have capabilities for facilitating these steps.

- A surface energy-balance method (SEBAL) was used for mapping the distribution of the water flux leaving the basin, i.e. evapotranspiration (ET). It is a physically-based (deterministic) model, which uses the relation between the surface physical properties and the electromagnetic energy recorded in the satellite images.

Outputs of the two types of analysis were used to define the water demands of the different sorts of users in a wet (1993/1994) and a dry (1999/2000) hydrological year. The demands calculated were then stored in the HYMOS database.
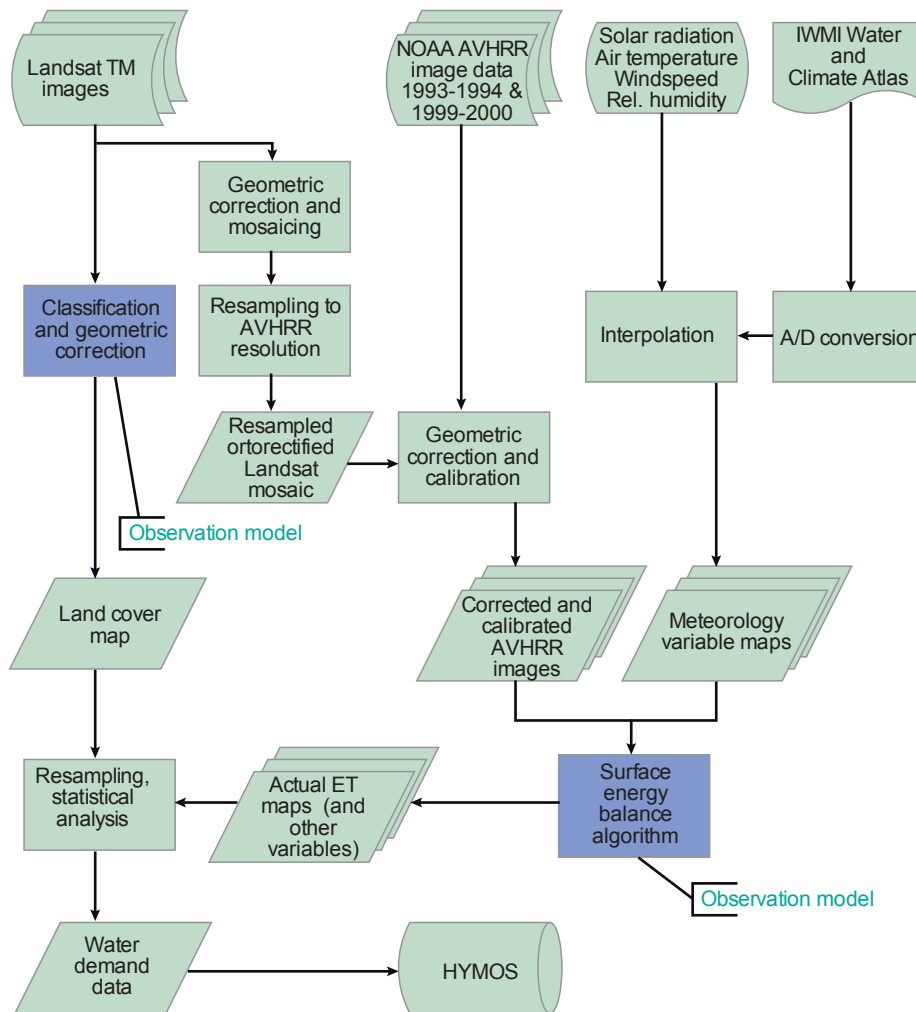
11.9.6 Process models

Allocation of water to different users requires a careful evaluation of the water resources available versus the water demands and priorities. The process modelling part of the DSS actually comprised two models:

- The RIBASIM (RIver BAsin SIMulation, developed by Delft Hydraulics) model, which distributes available water resources to users according to priorities and
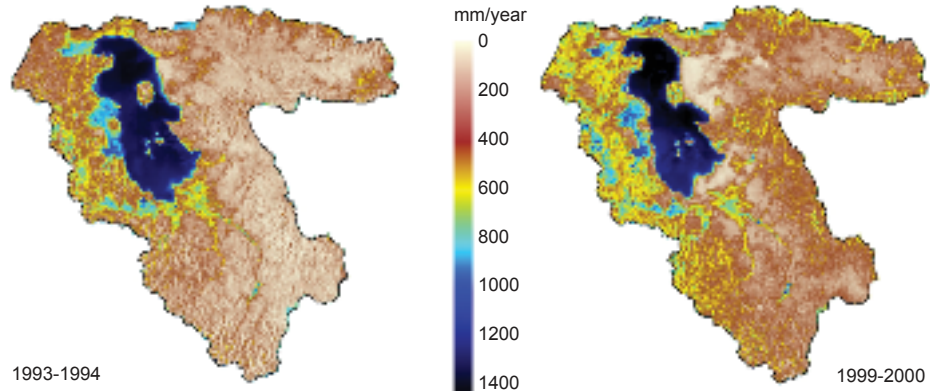
Figure 11.39
Example of the results of the surface energy-balance calculations: actual evapotranspiration maps of the two hydrological years analysed.
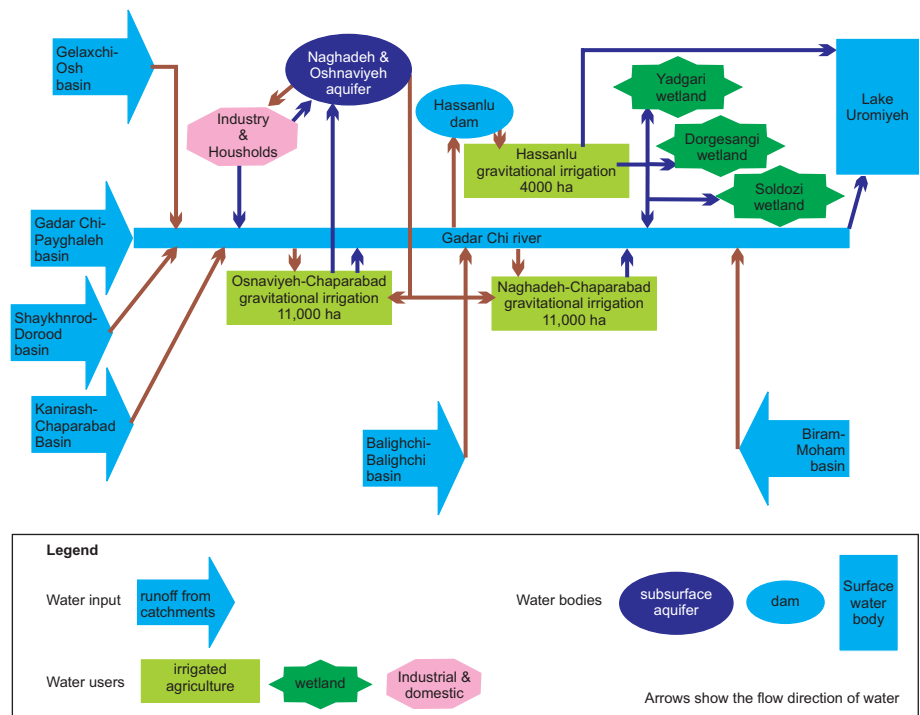
Figure 11.40
Schematic representation of water resources and users in the Ghadar Chai basin: the actual situation.

needs. It is able to simulate the flow of water from the headwaters of rivers to the lowest point of their basins—in our case, from the mountains to Lake Uromiyeh. Water flows can be traced through river channels, lakes, reservoirs and wetlands, as well as groundwater flows from sources to users. The logical structure of the actual situation in one sub-basin (the Ghadar Chai basin) is shown in Figure 11.40. Not only actual situations, but also future scenarios can be simulated with RIBASIM: Figure 11.41 includes four new irrigation schemes, a new dam and a new fish pond. RIBASIM can directly be linked to HYMOS. All the data integration steps are thus made outside the model.

• The DEFNITE (DEcisions on a FINITE Set of Alternatives) model, which is a tool for impact assessment and scenario analysis. A wide range of decision-support tools are included, e.g. multi-criteria methods, as well as tools for cost-benefit

and cost-effectiveness analyses, are available. For example, DEFNITE can compare different outputs (scenarios) from RIBASIM by assigning weights and assessing the most reasonable option. In this sense, DEFINITE is not a process model itself, but a tool to characterize and analyse the results of process models.

Several scenarios with different priorities and different levels of irrigation development were calculated. The DSS proved that actual water use is, unfortunately, not sustainable. This result poses an enormous challenge for the water managers of the region, especially since demand for water is still rising as a result of increasing social and economic pressures.

11.9.7 Conclusions

Data of different kinds, obtained from several sources, were integrated in the decision-support system developed for the Lake Uromiyeh Environmental Management Plan. The complexity of the situation required various methods of spatial and temporal re-sampling to create a data set that fitted the analytic methods proposed. Statistical and deterministic observation models were used to calculate the input data for a process model, which provided water allocation scenarios in all the sub-catchments of the Lake Uromiyeh basin. The DSS only supports decisions; it does not make decisions. The modelling results of the study proved that there was a need for new management plans that would match available resources with demand in a sustainable manner.

Acknowledgements

The Lake Uromiyeh project was carried out with the support of the governments of Iran and the Netherlands. The main Iranian contributors were the Water Research In-

stitute; the water, environmental and agricultural authorities in the provinces of East and West Azerbaijan; and Yekom and Pandam consulting companies. Dutch partners included ITC, Deft Hydraulics and Water Watch. As international consultants, Wetlands International and Dr Mike Moser participated in the project.

Chapter 12

Use and Users

Corné van Elzakker
Yola Georgiadou
Thomas Groen
Norman Kerle
Joan Looijen
Andrew Skidmore
Richard Sliuzas
Alexey Voinov
Eduard Westinga

Introduction

Geoinformation systems and information products need to be adjusted to their uses and users. This can be considered as a design process to which we can apply a systematic approach (see Figure 12.1). Who exactly are *users* of spatial information? One view could be that users are those who use a system without the complete technical expertise required to fully understand that system. As most GIS and EO applications are complex, and since almost all maps today are produced by some combination of GISs and EO methods, by this definition virtually anyone who has ever looked at a map is a user: there will be components of the hardware, software, and management or data systems that even an expert is unlikely to fully understand.

At the same time, it would be wrong to think of a user as somebody who sits at the end of the research chain and is only fed information from various flows of observed or derived data. After all, as a recipient of spatial information, the user could have an important role in defining what information should be generated, as well as in what form it should be presented. Moreover, it is often difficult to distinguish the producer of information from the consumer of that information. Perhaps the term stakeholder, which has also been used in the discussion on governance in Chapter 1, is more appropriate, as it connects the use of spatial data and information to an identified issue for which access to and use of spatial data and information are considered to be relevant and important.

It is clear that enormous volumes of data are being generated. This phenomena was identified by the editors of a special 125th anniversary issue of the prestigious science journal *Science* entitled "What we don't know" as a significant scientific challenge.

Among other things, the editors posed 25 key scientific questions, one of which was "How will big pictures emerge from a sea of biological data?" [90]. Such a question can be easily broadened to "How will big pictures emerge from a sea of data"!
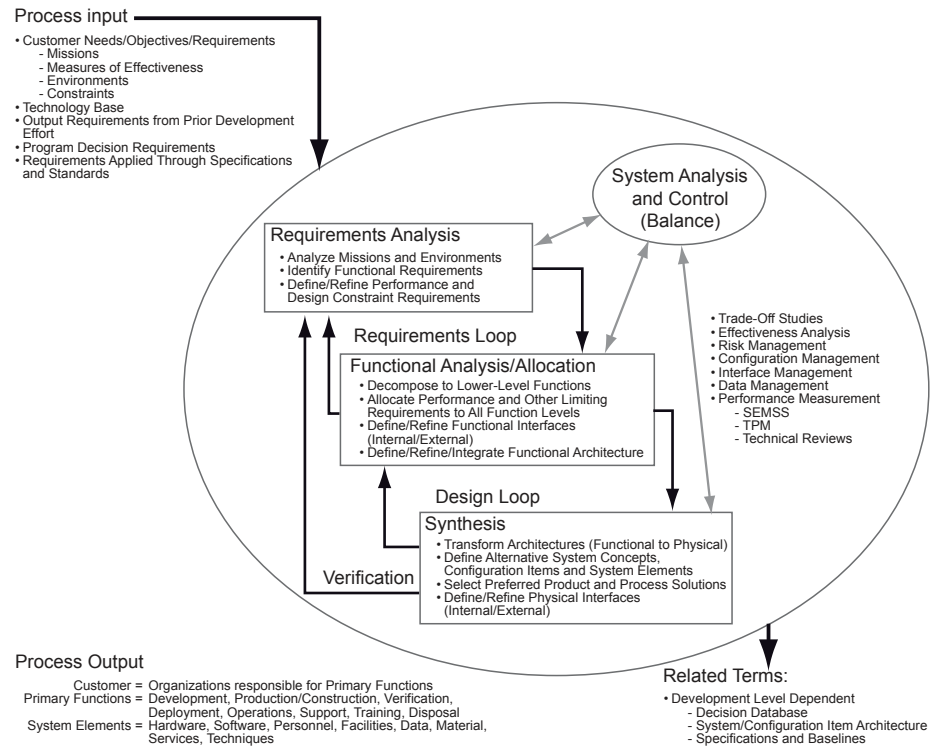
Figure 12.1
Systematic approach to the design of a new product for the consumer market

The objective of this chapter is to illustrate the enormous variety in applications of spatial information in daily life. Each application is of course derived from people perceiving a lack of spatial information related to a specific problem. After reading this chapter you will be able to explain a number of problems for which the solutions require spatial information. You also will be able to identify the stakeholders involved and their information needs. Each section in this chapter on *Uses and users* will, therefore, define the problem, describe characteristics of stakeholders involved and discuss responses or, where applicable, possible solutions. It is important to note that the usability of any system's products is determined by how practical and convenient the capability of that system is for users. In other words, a specific geoscience application becomes truly useful only when it provides a solution to a broad societal problem (such as a navigational aid, as described in one of the following sections, helps us optimizing our travel between locations given constraints such as road and traffic conditions.

There is a great deal of geographically referenced data available that is generated by Earth Observation, as well as other sources such as censuses and field observations. These data need, firstly, to be grouped, analysed and processed in order to generate useful information. In other words, to provide answers to *how*, *what*, *where* and *when* questions. The application of data and information allows us to answer *how* questions—for example, we can show *how* a remotely measured indicator such as the NDVI (Chapter 9) is related to green vegetation biomass (i.e. a linear relationship that saturates with an asymptote at higher levels of green vegetation biomass). The next

level is an understanding of *why* a phenomenon occurs—for example, *why* NDVI is linearly related to green vegetation biomass. *Where* and *when* questions are of course the obvious ones in a GI Science context. Finally, we can use this understanding to come to a "wise" conclusion, such as predicting food production in an area and, if a food shortage threatens, to carefully consider options for short-term and long-term alleviation of that shortage. Such information products are therefore important inputs for the governance of famine relief efforts.

Applications may vary in their level of complexity. Multiple users may access geoinformation generated for applications as diverse as the assessment of flooding hazards, monitoring the condition of coastal defences, or managing nature conservation areas. Sometimes access is structured via GIS interfaces, specifically designed for multiple users. These issues are discussed in some detail in Sections 12.2-12.4. In Section 12.5 a further level of complexity is illustrated with the description of a flexible geoinformation application designed to allow multiple users in the Netherlands who are interested in spatial planning to exchange digital spatial plans at a national level.

12.1 The users of route planning and navigation systems

12.1.1 Introduction to route planning and navigation

For a number of years, computerized GISs were expensive and operators had to possess a high level of skill to be able to carry out even basic tasks. These systems were, therefore, inaccessible to all but a limited few. It was through one particular application, however, that the public at large also became familiar with GISs: route-planning systems. These systems, which were originally distributed on CD-ROMs, but are now widely available on the World Wide Web, help users to find the best answer to a fundamental question: "How to get from A to B?". These users may be individuals who want to move in geographic space for all kinds of reasons and by various means of transportation (e.g. car, bicycle or on foot), but they may also be professional transporters for whom time and distance have economic consequences.

Road networks are physical geographical structures and network analysis is the GIS operation applied for route planning. But route planning is only one aspect of answering the question "How to get from A to B?". Once the route has been determined, users will actually have to follow it. In the past this was usually done by consulting printed road maps and checking road signs, but modern navigation systems such as in Figure 12.2 assist users while moving from A to B by giving visual and/or verbal instructions such as "turn right after 300 m".

Figure 12.2
An example of a car navigation system.

This section presents the use and user requirements related to route planning and navigation. These requirements are the starting point for a discussion of the main components of this particular kind of GIS. The discussion will not be a technical one (for that, we refer readers to other chapters in this book). It will rather be approached from the perspective of system's use and users. The section concludes by giving us a glimpse of the future in relation to personal navigation systems.

12.1.2 Use and user requirements of navigation systems

Throughout the course of civilization, the need for travel and transport information has increased enormously. Fresh flowers are grown in Africa and transported to and sold in the Netherlands. People no longer necessarily live in the same town as where they work or go to university. And we take holidays in increasingly remote places. This has been made possible by improved means of transport, means that can move goods and people through space relatively fast and cheaply. In answering the question

"How to get from A to B?" a user aims at effective, efficient and satisfactory relocation in space, whereby she or he will normally make use of existing physical structures on (or above or below) the surface of the Earth. These physical structures consist of roads and paths. In this context, the *how* in the question means "along which route?" and not by which means of transportation, although the latter is, of course, a very relevant question as well. In view of the solution to be provided, it really matters whether the transport takes place by car, truck, airplane, bicycle, subway, train, boat or on horseback or foot.

There may be great differences in the purpose of relocation: transportation companies want to move goods or people from A to B in the fastest and cheapest way, with the lowest fuel costs. If a user wants to travel from A to B by truck or car, this may not necessarily be done by the shortest path. On the other hand, people on holiday, or those touring on a free Sunday afternoon, may not be interested so much in speed or even distance. They may be interested more in scenic and attractive routes with not too much traffic. In many respects, therefore, user demands and characteristics differ greatly. What they all have in common is that they do not want to get lost.

In their turn, governments also have an interest in optimal route planning and navigation, because they want to facilitate efficient travel and transportation. The latter, in particular, may lead to traffic congestion, the need for more roads, and unnecessary air pollution. In special cases, governments will want to prevent risks by not allowing the transport of hazardous substances in the vicinity of centres of high population concentration.

All in all, there is a need for tools or systems that help people to answer the question "How to get from A to B?". This question has two aspects, both of which can be related to the specific use and user requirements: (a) optimal route planning and (b) its navigation, i.e. actually following the planned route. In the past, it was very much up to the individual to find answers for these two aspects. People used their own experience, or the geographical knowledge of others, or they used traditional road maps. It is not hard to imagine that this often led to unwanted detours and unnecessary waste of time, money and fuel, not to mention marital disputes in the car! Thanks to the route-planning and navigation systems available today, the situation has clearly improved.

12.1.3 Navigation systems from past to present

In the early days of route-planning applications, a user had to buy a CD-ROM containing the application and the data in a shop and insert them into a personal computer. The resulting route description in words (containing distances, directions, locations and, possibly, time indications) and/or route maps could be printed on paper for navigating in the car. So, users would need a PC and printer. Nowadays, route-planning applications are freely available on the World Wide Web. Routes may well be planned with, for instance, Google Maps, one of the many route-planning applications now available on the Web.

In the past, navigation focused on answering the question "where am I?" and on taking subsequent decisions on directions to follow, turns to take, etc. The answers and decisions were based on linking map displays and route descriptions with what people saw in reality (including road signs). In the development of navigation systems, the biggest problem to be solved was that of assisting users to determine their (or their vehicle's) geographical position. In this respect, the availability to the public of GPS (Global Positioning System) after 2000 gave an enormous boost to non-military navigation systems. These systems now come as standard features in mobile phones and cars as well as stand-alone navigation devices.
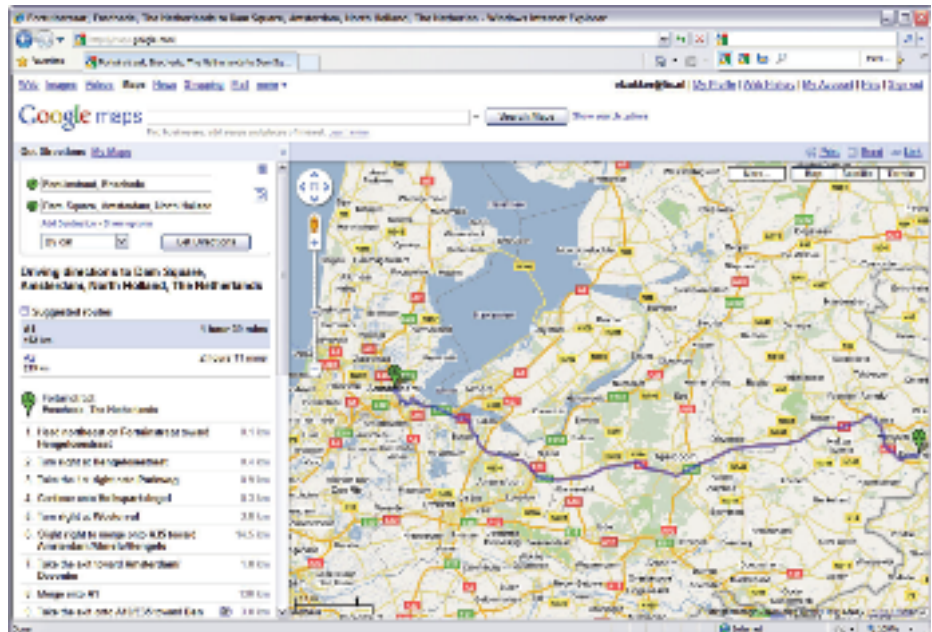
Figure 12.3
A route planned for a motor vehicle with the help of Google Maps going from ITC, starting at the Fortuinstraat (exit of parking lot) and ending on the dam square in Amsterdam. Source: [40]

The systems are not large (typically smaller than the palm of one's hand), primarily consisting of a display—a (touch) screen—on which the route map and further instructions are shown. With navigation devices such as those manufactured by TomTom or Garmin, users can also plan routes by typing in a destination; a point of origin can also be typed in, unless you are starting from your current position, which the system already knows via the GPS receiver or other means of positioning that is part of the navigation device. After the user accepts the proposed route, the device will help him or her to follow the route by giving verbal instructions or indicating directions to be taken by graphic symbols such as arrows. Of course, the system must also have sufficient capacity to store all the geographic (digital map) data of the area in which the user is moving. The required geographical data may be stored on, for instance, micro SD cards, or may be obtained through mobile Internet.

Geographic data are an essential part of systems for route planning and navigation. There are companies that specialize in collecting these data (e.g. TeleAtlas and NavTech), which they then sell to other companies that build and market navigation systems. The data that are collected are essentially data about physical structures, consisting of nodes and lines (edges) that connect these nodes. The lines may be road segments or, for instance, bicycle tracks or footpaths. These physical structures do not only have geometric properties: all kinds of attributes can also be attached to the line segments. Examples of such attributes would be the average driving speed on a particular road segment (calculated on the basis of speed limits, nature of the road, the presence of traffic lights and even data about the amount of traffic), its scenic attractiveness, the surface of the road and whether cycling is allowed on a particular road stretches. The more attributes there are, the better the various uses and user requirements can be met.

In addition to the road (line) segments, the geographic database will also contain many nodes and their (x, y) locations. These nodes are potential destinations of users, e.g. Points of Interest (POIs) such as a railway station or a soccer stadium, which may also be useful as navigational features ("after passing the church on your right-hand side,

Reality through
the eyes

Perception/cognition of
reality inside the mind

Where am I??

Representation of
reality on
mobile screen

User of
geo-mobile application

Figure 12.4
Personal geo-identification
through the interaction of
three different information
sources: reality, cartographic
representations of reality, and
individual cognitive maps in
the user's mind. Source: [26].

turn left"). Potential destinations in an address database may also be stored as points on a line segment.

Clearly, the size of a geographical database required for a route-planning or navigation system is enormous. After all, such a database would have to contain a comprehensive collection of geographic names (e.g. street, town and city names). Some databases may even contain several versions of the same toponym, to cater for differences in spelling.

The quality of the physical-structure data is crucial for determining whether users will trust and appreciate a route-planning and navigation system. In this context, the accuracy and currency of the data is also important. In 2009, a navigation system for inland waterways had to be taken off the market because of mistakes in the depths stored in the database!

Knowledge of the geographical position of a user and his or her system is an essential requirement for the navigation functionality of any particular system. The earliest in-car navigation systems determined the position of the car by means of transmitters and receivers mounted in the vehicle and on lamp posts along the streets, in combination with mathematical calculations related to the steering movements of the vehicle and distances covered from the starting point. Obviously this kind of positioning was rather cumbersome and prone to error. A real break-through occurred in the year 2000, when a dedicated military satellite-based positioning system became available for civilian use: the accuracy of the system was no longer reduced for military reasons. In the course of the last decade, the quality of GPS receivers in navigation systems has increased enormously, so the movements of the users can be followed dynamically with sufficient accuracy.

The software of route-planning and navigation systems usually displays the geographic data from the database in map form. The user may determine which area is retrieved by means of entering addresses or geographic names. Users may also interact with the map display and pan to the area they are interested in. Zooming in and out of the map display is also a very common feature of interacting with the system. When the system contains a GPS receiver, the location is usually represented on the map display and the surrounding geographic data are automatically shown. During navigation,

the GPS location symbol moves smoothly over the road symbols because the software *snaps* the incoming GPS coordinates with those of the road networks in the database.

When users want to plan a route between a particular point of origin and a desired (indicated) destination, they will start executing a real GIS operation on the geographic database. This operation is called network analysis (see Section 9.5) and the typical function called upon is optimal path finding. What is optimal depends on the user as well as being related to the attributes that are attached to the line segments in the database. Users may opt, for instance, for the shortest path, or the quickest or most scenic path. They may also indicate whether they will travel by car, bicycle or on foot. As, for instance, the direction of travel is also important (particularly in the case of one-way streets), you may imagine that the routing is a far from simple mathematical exercise.

Once planned, a route is stored in the system and the user is guided along it. Navigation instructions are automatically provided when needed, i.e. when certain locations have been reached (determined from the incoming GPS data). If users deviate from the planned route (consciously or otherwise), the system will first give a warning and suggest that the user returns to the planned route or will later automatically re-calculate an alternative route to the destination.

Compared to the traditional printed maps, however, the roles of map displays in these systems have changed. These maps can no longer be regarded as databases that show all geographic information that may be relevant for route planning and navigation. They now provide a dynamic and interactive interface to the system and the geographic information it contains, and they present an overview of the planned route (for cognitive confirmation) and supporting navigation, particularly in complex geographical situations. Map displays are thus specifically designed to meet different user requirements. A clear example of this is, for instance, the use of an oblique map view during navigation and the rotation of the map view in relation to the direction of travel. System/map designers also deal with cartographic generalization (e.g. leaving things out, or representing them in a simplified way), brought about by the limited size of display screens and the frequent zooming out (and in) by users. Such generalization is steered by the characteristics and requirements of the users.

Route planning and navigation in practice

René Boensma works for Velda, a company in Enschede, the Netherlands. Velda assembles and manufactures various products, such as water filters and pumps, for the maintenance and care of garden ponds. These products have to be shipped to retail sellers such as garden centres throughout the Netherlands and other countries in Europe. René used to work for the company as lorry driver, but nowadays he is more involved in distribution planning. René was interviewed together with the company's Assistant Director, Alexander Dalenoort.

The distribution of Velda's products to its customers has two aspects: route planning and the actual transport (including lorry navigation) of the goods. Both are subject to the economic maxim "time is money". In this commercial setting it is not just a matter of finding an answer to the question "How to get from the company to the customer in the cheapest/fastest way?" After all, the company has more than one customer and it is inefficient to travel to a customer with a half-empty lorry. In the planning of a lorry's trip several factors have to be taken into account: the orders, which are placed in sequence of date of delivery/urgency; the geographic location of customers; the volumes of the orders and the carrying capacity of the lorry; the height of the lorry (in view of the height limits of viaducts and fly-overs); and much more. In the past, René did this planning mainly on the basis of his experience with travel times, as well

as with the help of printed wall maps provided by office suppliers. As a lorry driver, René trusted his experience, road signs and detailed printed maps such as town plans to navigate his route.


René Boensma, a route planner and former lorry driver of the company Velda (photo courtesy Velda BV).

Since the year 2000, transport companies use computerized route-planning and navigation systems. At the same time, traffic has become much more congested and route-planning and distribution problem much more complicated. For instance, customers demand specific delivery times (e.g. before 10:00 a.m.), and garden centres, which have united into big purchasing organizations with a network of distribution centres, have contracted specific transportation companies for the collection and delivery of goods. A relatively small company like Velda, with only five lorries, could not make deals with its customers anymore without accepting the condition that the goods had to be transported by a specific trucking company. These big companies also transport the products of other suppliers to garden centres and this may simplify the route-planning problem if the complete carrying-capacity of the lorry is taken up by goods for just one customer. On the other hand, the factors mentioned above and the increasing importance of the "time is money" motto have led to the development of more sophisticated route-planning software and even "fleet management" systems that only big transportation companies can afford. In fleet management systems, the actual real-time position of lorries is always known, allowing the central office to take measures in the case of problems (e.g. call the customer if there is a delay or re-direct a driver).

For these reasons, despite the overall growth of the company, Velda disposed of four of its lorries and put out to contract the transportation of its products to third parties. There is now only one lorry left and its driver definitely profits from the possibilities offered by modern traffic navigation systems. However, both René and Alexander say that they would never go anywhere without a traditional, printed map. Such a map gives a better overview (before and during the trip) and can be used in cases of emergency (loss of GPS signal or a power problem with the navigation system). Also, because of their experience and knowledge of local situations, they sometimes do not trust the route indicated by the system. René Boensma concluded by saying that for an experienced person like him, the map display on the screen of a navigation device plays a very important role in his decision-making on the road, as well as confirming his local knowledge.

Another example of a situation in which careful routing is required is when transporting hazardous materials such as explosives, flammable liquids, oxidizing substances, poisonous gases and radioactive materials.

Figure 12.5
Optimal routing for
transportation of petrol
derivates in Lalitpur, Nepal,
based on sample weighting
scenario for the objectives
involved [5].
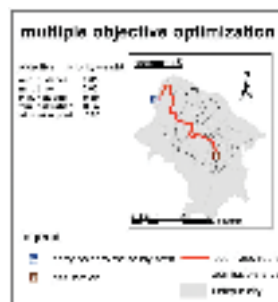
Many factors play a role in the process of route optimization for the transport of hazardous materials, such as the number of people exposed to the risk (e.g. schools along transport routes), the accident rate on certain route segments, and travel times and distances. In many countries, most weight is given to economic factors and the accompanying risks are somewhat neglected. This can in part be explained by difficulties in finding suitable solutions for the route optimization problem. In response to this challenge, Avendano Castillo [5] developed a route-optimization model that can be used as a decision-support tool to take into account economic and risk factors when assessing routes for the transport of hazardous materials. As a case study, he selected the transport of petrol derivates in the city of Lalitpur, Nepal. A technique called "multiple-objective mathematical programming" was applied to calculate optimal routes that were dependant on five objective functions that were to be optimized (minimized): travel time, travel distance, risk for the population, risk for the urban environment, and risk related to natural hazards. Figure 12.5 shows an example of optimal route obtained with that model.

12.1.4 Future developments

The usability of the systems is still far from optimal and many improvements can be made. The attention of users is often diverted by the information on the display screen and by the verbal instructions. User research is required, for instance, on the relationship between geographic reality, the representation of reality by the system (through map displays, but also through verbal and graphic navigation instructions) and the mental maps in the minds of the users. Questions like "where am I?" and "which direction should I take next?" should and can be answered better with the help of route-planning and navigation systems. In this context, landmarks play an important role. In addition, the spatial awareness of people should increase, if only to prevent that they become totally lost if there is no longer any GPS signal, or when the device's battery is exhausted. Some spatial awareness is thus required in order to prevent that users blindly follow the navigation instructions of the system. Such behaviour has resulted in car drivers ending up in canals at night and truck drivers getting stuck in narrow village streets. In other words, systems should not only be based on use and user requirements, but should also foster and develop people's spatial abilities.

A problem that still awaits a solution is that of positioning indoors (where satellites are not "visible" for the receivers). Such indoor positioning is relevant for navigation in, for instance, commercial shopping malls, museums and subway systems. At the moment, there are simply no accurate, low cost, small-sized and infrastructure-free positioning solutions available, but much research is being done in this field. Avenues being explored include dead-reckoning systems (inertial systems with, for example, accelerometers, gyroscopes and magnetometers) and reference-based systems (such as cellular mobile-phone networks, WLAN/Wi-Fi, WPAN/Bluetooth, RFID, laser, ul-

trasound, Ultra-Wide band).

One interesting development is the collection and use of temporal geographic data for route planning and navigation. In fact, it would be useful if the amount of traffic on various road segments (which influences traffic speed) at different hours of the day and under different weather conditions could be used for dynamic route planning (e.g. calculate alternative routes to avoid traffic jams). Such data are already available, either as *live* data or as averages from the past, and their implementation is already operational. The data are collected by sensors in the surface of the road, or through other means. TeleAtlas, for instance, has an agreement with telecom provider Vodafone to get data about the lines of cell phones standing still on a particular road segment. This is an example of users themselves generating data that may be used by other users.

Other developments in this field, known as neo-geography, also affect systems for route planning and navigation. In the OpenStreet Map initiative, for instance, it is road users themselves who collect data about the road network, not only data about the geometry of these roads (simply stored by recording GPS traces), but also about their attributes. In a similar initiative in the Netherlands, cyclists (volunteers) collect all kinds of attributes of cycle paths (such as safety at night and the nature of the path's surface) that would never be collected by a commercial company. These data are made available to other users for route planning and navigation.

12.2 The users of early warning systems

12.2.1 Introduction to early warning systems

The WHO Collaborating Centre for Research on the Epidemiology of Disasters (CRED) defines natural disasters as events that either kill at least 10 people, affect at least 100 others and lead to a state of emergency being declared, or a call for international assistance to be given [24]. The annual occurrence of natural disasters continues to rise globally. And as vulnerability to disasters is also rising worldwide, the economic damage incurred and the number of people affected will continue to increase.

Disasters constitute a severe impediment to economic growth. This especially holds for economically less-developed countries: they have suffered more than 90% of all disaster-related fatalities and have been disproportionately burdened by the economic losses. Since around the year 2000, between 500 and 600 natural disasters have occurred annually somewhere in the world and these events routinely lead to global economic damage that is estimated to cost more than US$ 100 billion per year [55, 24].

The natural disaster that affects most people is flooding. In China, for example, by some estimates up to 200 million people were affected by flood events in 2007 alone. Most developing countries regularly experience some form of flooding, be it the result of storm surges along coastlines, the consequences of their annual rainy season, or due to snow melting in the mountains. There has also been a strong debate about hazards—which may lead to disasters—becoming more frequent and severe as a result of global warming. The changes we have already witnessed and that are projected to occur, affect various aspects of our environmental system. Those changes relate especially to general shifts in precipitation regimes that can either lead to stronger flooding or more severe droughts. In addition, countries such as Bangladesh, but also the Netherlands, have to cope with land subsidence, making coastal protection more challenging.

We can, therefore, expect especially flood-related disasters to get worse in the future. This trend is further intensified by increasing urbanization, as most urban areas are located close to the coast or major rivers [53]. Another factor is the growing global population. The poorer members of populations are increasingly marginalized. This forces them to settle in hazardous areas such as unstable slopes or flood plains, where even modest flooding could lead to a disaster. Such spatial marginalization is usually also coupled with socio-political marginalization, meaning that lack of access to education, adequate healthcare and sanitation, but also lack of access to financial resources and political (lobbying) power, lead to increased vulnerability [30].

Thus, there are many good reasons for studying flood-related hazards and working towards flood-disaster mitigation. This section describes how our GI science tools relate to these hazards, the users of such tools and data or the information they ultimately deliver. Two individuals are introduced who are not only studying flood-related hazards but who are actively involved in keeping people safe from flood events. We will learn more about how they do that and why they are frequent users of GI science.

12.2.2 Users of early warning systems

Asking who uses GI science on flooding leads to a surprisingly broad answer. In principle, anyone who has ever been affected by flooding or who has ever sought early warning information or advice on how best to cope with or recover from flooding is a potential user of GI Science. The information products do not have to be very sophisticated: anyone watching a weather forecast to learn about threatening rainfall situations or looking at a flood evacuation plan fits this description. At another level of

sophistication, there are professionals who actually use geodata and tools to prepare those forecasts and plans.

We can, therefore, distinguish between expert users who are equipped to use raw data and sophisticated tools to create relevant information and more casual users—laymen—of this information. The latter group of users has little understanding of the models or data used to create the products intended for them. The sophistication of the professional group strongly depends on the technological means available, the funding needed to acquire the data and tools, and the technical capacity of staff. It is easy to see that there can be great differences in how flood hazard is being dealt with, especially between more and less developed countries. There are also comparable differences amongst lay users. In western countries it is normal for people to have internet access at home. They have ready access to large amounts of hazard-related information, as diverse as actual information on river levels, detailed local storm surge warnings or real-time rainfall radar maps. People in poorer countries, on the other hand, may only get some radio reports or information via community bulletin boards. The rest of Section 12.2 will focus on flood risk management to illustrate the use of geoinformation in early warning systems.

Olaf Neussner, a flood-risk specialist from German Technical Cooperation (GTZ) working in the Philippines.

Interview highlights

"Owing to limited availability and/or reliability of official geoinformation in the Philippines, we rely to a considerable extent on gathering data ourselves."

"Without maps, and the remote sensing and GISs to produce them, it would be very difficult to establish the Flood Early Warning System".

"…we would be happy if the official mapping agency would update the base map of the area (50 years old), and share their GIS file".

Olaf Neussner has been working in the Philippines for over 14 years and, together with national and regional organizations, has devoted much effort to flood-risk management. In his own words, he works towards reducing casualties and damage caused by river flooding in Region VIII of the Philippines (Leyte and Samar Islands). He explains: "We work on Flood Early Warning Systems (FEWS), and assist local governmental offices in the establishment of such FEWS. This includes the identification of

flood-prone areas with GI Science tools". Olaf and his colleagues make use of EO data and geodata processing methods for hazard, vulnerability and risk assessment, as well as the actual forecasting of floods. They also integrate administrative data obtained from publicly accessible sources and from governmental agencies, such as maps of flood-prone areas provided by the Mines and Geoscience Bureau. Valuable input also comes from people in potentially affected areas themselves, thus bringing together the professional and lay users of GI Science.

The approach applied in the Philippines is similar to what is being done in other countries, although variable approaches to flood-risk management exist, as do solutions at different scales. The Netherlands has suffered greatly from major flood events in the past. After all, it is a country with a long coastline and in which some 15 million people live in areas that lie below sea level. Moreover, the Rhine and Waal rivers run through the Netherlands and may bring flood waters from neighbouring Germany.

12.2.3 Historical flood events

Europe's worst flood event in recent history occurred during a storm flood in the North Sea on the night of 31 January 1953, affecting mainly countries on the edge of the North Sea. The tidal surge exceeded 5.6 m above mean sea level, with the resulting flood and storm-driven waves spilling over and breaking through coastal defences. Fatalities included 1835 people in the Netherlands, 307 in England and 28 in Belgium; 150,000 ha, or nearly two entire provinces, were submerged. A photograph of the flood is shown in Figure 12.6.

Figure 12.6
Netherland's worst flooding in February 1953.

A more recent flood occurred along the Rhine/Waal rivers in 1995. The death toll was low (four people), but some 250,000 ha had to be evacuated and several billion Euros in damage was suffered. The Netherlands learnt valuable lessons from these events. An enormous engineering project, called the Delta Works, was initiated only days after the 1953 flood and a carefully maintained system of dikes to protect coastal areas has been set up. Figure 12.7 shows one of the massive barriers near Rotterdam that can be closed when a storm surge is forecast, principally to protect the areas around the mouths of the Rhine, Meuse and Schelde rivers.

Bas Overmars works at the Provincial Office of the Province of Gelderland, the region in the Netherlands that faces the highest risk of Rhine/Waal river flooding. He ex-

plained that in general the river banks are heavily fortified with dikes (some of the 2500 km of dikes protecting the country border these rivers) and drainage systems. Still, flood management relies strongly on flood-model predictions and real-time data input such as precipitation and river gauge data from Germany. It also includes carefully planned evacuation scenarios, multi-agency cooperation plans and emergency drills. These are supported by effective communication channels, not least with the population that has the highest potential of being affected. Bas Overmars, highlights what the main purpose is of his flood-risk assessment and management work. It mainly involves acquiring knowledge of the water-cycle system and improving the availability of information for disaster mitigation during periods of high water levels in Gelderland (the Netherlands) and North-Rhine Westfalia (Germany). An important part of the programme is the organization of flood-response exercises. This ensures that the knowledge acquired is maintained within an organization. He emphasizes that it is important to describe such arrangements in written agreements.

12.2.4 The GI Science behind flood risk management

Both the Philippines and the Dutch example show that a number of types of geodata and tools are important in flood-risk assessment and risk management. Let us look at each of these in turn. What exactly is risk? Well, simply put, it is the resulting product of hazard and vulnerability. This means that we have a risk when a hazard, such as flooding of a certain magnitude and a certain periodicity, intersects spatially with what are called elements at risk (EaR). Such EaRs include people, but also their homes, their cars, the infrastructure they use and their industry; in short, anything that is of value and can suffer damage or destruction by the type of hazard under consideration. Thus the risk is high where we have a high concentration of high-value EaRs (e.g. major cities) and a large hazard. Note that this can result in a place that suffers from limited flooding several times per year having a level of risk that is similar to that of a place that can be hit by a rare but major flood.

To say anything about risk we must, therefore, have good knowledge of flood hazard. Such knowledge can be derived from historic records of flood events. Where we have a seamless record of flood parameters such as frequency, time of onset (i.e. speed at which the flood waters rise) and depth and duration of past floods, we can clearly establish the nature of the flood hazard we face, provided that the causes of the flooding have not changed significantly. This may not be the case everywhere,

particularly in view of the climate-change related modifications to our natural systems. In addition, people have a tendency to straighten and dam rivers, clear forests that may otherwise absorb rainfall or slow down surface runoff, and seal the ground (e.g. pavements, buildings, parking lots), thus preventing water infiltration. This may lead to hazardous situations that change quite dramatically over time. A better way of understanding current flood hazards might be to use modelling techniques.
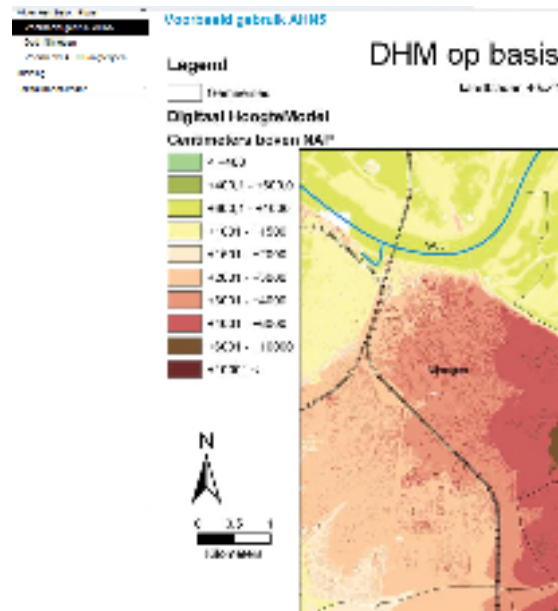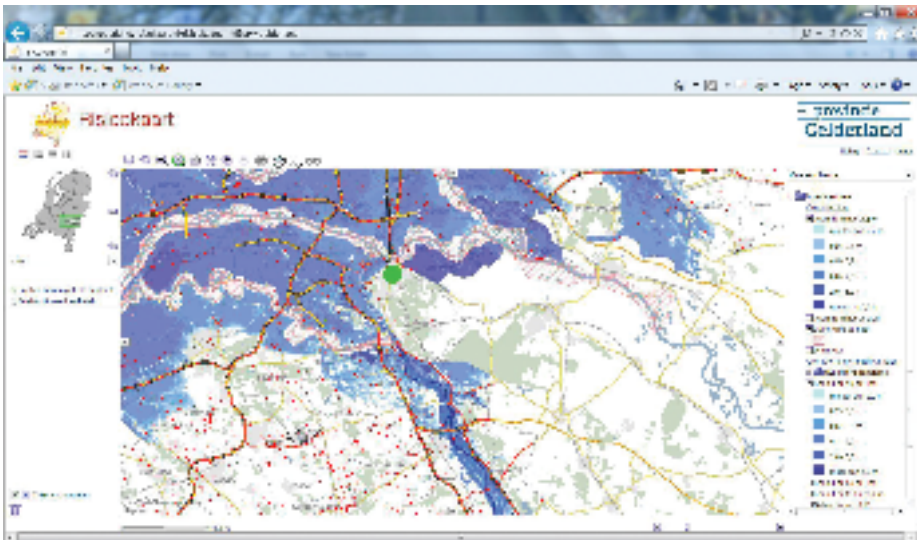

Figure 12.8
Digital surface model of the area around Nijmegen, the Netherlands. Source: [38].

Characterization and quantification of floods can be done with several hydrological and hydraulic models. For the Netherlands, a basin model can be used to estimate the discharge in a river as a function of rain and basin characteristics, e.g. shape, size, river cross-section, land cover and soils. Such a model produces at predefined locations along rivers hydrographs for rainstorms of a variety of periodicities (e.g. annual or 100-year floods). However, on flat and complex terrain, such as that of the Netherlands, this is not enough, as rising water will leave its river channel and spread out over larger areas. To understand such a situation, water flow over a dry surface in 2D space is modelled. To be able do this, a good digital representation of the terrain (i.e. DSM introduced in Section 4.5) is needed, one that includes above-surface features such as buildings, around which water would typically flow. An example of such a DSM is shown in Figure 12.8.

In the Netherlands, GI scientists are in the fortunate position of having a highly accurate DSM derived from airborne laser scanning (LIDAR) data that serves as input for the model. Laser scanning is ideally suited because it can obtain complex elevation information that allows the detection of both the top surfaces (tree tops, roofs, etc.) and the ground surface beneath vegetation. With appropriate filtering, a true representation of the ground elevation (DTM, see Section 4.5) can be derived. Or one can create an elevation model that retains all solid flood obstacles (buildings and other infrastructure) but has vegetation removed (since vegetation does not form an impenetrable barrier to surface flow). LIDAR's multi-elevation mapping approach becomes really useful where important structures, such as dikes, lie beneath vegetation. In the Netherlands such sophisticated modelling is already being done. Users of such a LIDAR-derived DSM do not necessarily have to be laser scanning experts. This is

shown in a product that is developed by the Dutch provinces: an interactive risk map (see Figure 12.9). This map includes various data layers that the user can switch on or off. Also the extent of the map can be determined by the user by zooming in or out. Anyone interested in finding out the risk for natural disasters such as flooding can access the maps. Also other types of disasters e.g. human induced disasters due to storage or transport of chemicals can be mapped in this way by laymen.
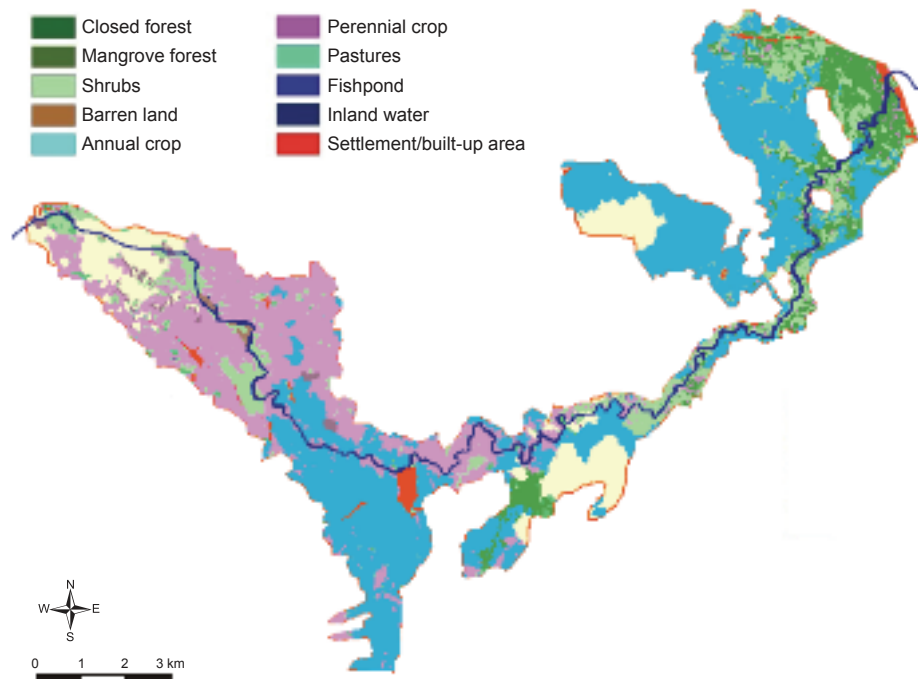

Figure 12.9
Flood depth in case there were no dikes. Flood levels presented are based on extremes that statistically occur once in ten years. The green dot indicates the city centre of Nijmegen. Only the Dutch territory is mapped. Source: `http://www.risicokaart.nl/`

Unfortunately LIDAR data is not yet acquired in all countries in the world. In the Philippines, for example, Olaf Neussner has to use other approaches. Owing to a lack of critical data and models, he cannot model flood scenarios. Therefore Olaf uses historical assessments of past floods involving observations from helicopter overflights or satellite images sensed during floods. High-water marks on buildings and existing flood-hazard maps are also useful input. This approach may be less accurate than the results of flood modelling, but it still provides a good approximation of the flooding a certain rainfall event can cause. Since Olaf also uses accounts from people who experienced a particular flood, it is more a participatory approach in which the latter lay users of flood-hazard information actually help professional users to create this information.

There are also historical sources of elevation information. Photogrammetric processing of aerial photos has provided accurate surface data. Where good topographic maps are available, an interpolation of digitized contour lines can provide digital elevation information. It does, however, not provide information on structures that have the potential to block water flow. Understanding and dealing with such potential data limitations is critical. For the areas in the Philippines for example, existing base maps are about 50 years old and existing hazard maps do not indicate the severity and probability of potential flood events.

To model flood hazard we also require land cover data, both to correct DSMs (e.g. to remove vegetation) and to extract surface roughness information. While topographic data can be considered to be relatively static, land cover data is more dynamic, especially in urban areas. Land cover data can be obtained from satellite image classification, as described in Section 6.2. In the Philippines, land cover information is often derived from medium-resolution satellite data (e.g. ASTER, SPOT). In addition, free information from Google Earth is also used.

Figure 12.10
Satellite-image-based land
cover classification of the
area identified as being at
risk of flooding by the
Binahaan river (Philippines)
(Source: Olaf Neussner,
GTZ).

Remote sensing data are useful to address the Elements at Risk (EaR). To quantify risk related to each EaR we need to know its type, spatial location, how much it is worth, and how susceptible it is to a hazard of a given magnitude. EaR detection can be done using high-resolution image data, such as from Ikonos, Quickbird or aerial photographs, and is useful for a detailed assessment in urban areas (for more information on the utility of remote sensing in disaster-risk management see [54, 131]. Determining the value of a given physical feature, such as a building, factory or bridge, usually requires some auxiliary data, as this parameter is almost impossible to extract from image data. For detailed work, structural engineers need to be involved to determine how a given structure will respond to a hazard, such as a flood of a certain height, duration and flow speed. In addition, there are also other EaRs, such as agricultural fields, which, if destroyed, have to be added to the list of damage. Neussner and his group also use satellite imagery to determine the types of agricultural use in areas susceptible to flooding (Figure 12.10). To estimate accurately the number of people who may be affected under a given flood scenario, census data are needed. An alternative is to map the number and possibly types of buildings and use an average occupancy rate, which may be calibrated with some field knowledge [68].

12.2.5   Monitoring as part of an early warning system

In general, river height downstream is a function of water accumulation in the upstream catchment, typically the result of rainfall. Other situations are also possible in case of storm surges or flooding resulting from excessive precipitation that exceeds the capacity of the drainage system in an urban area. To measure the amounts of rainfall, it is easiest and most accurate to employ a network of rain gauges. In the FEWS run by GTZ for the Binahaan watershed, for example, automated rainfall gauges are used as well as local observers [81]. Those observers read the manual gauges and transmit the data via a Short Message System (SMS) to the operations centre. They do that twice a

day under normal circumstances and more frequently during critical rain events. The same hybrid approach using manual observations and automated systems is used to monitor river water heights. The overall reliability of a fully automated system can be expected to be higher. However, the approach just described is necessitated by limited funds. Nevertheless, it also has the advantage that the people living in the watersheds affected feel a certain ownership of the system, as they are personally part of it.

In the Netherlands the situation is different. Flooding by major rivers is typically a consequence of large amounts of water accumulating in Germany, the result of precipitation or spring snow melt in Germany's south. The Rhine is a relatively large river, draining an area of some 220,000 km^2, about five times the entire area of the Netherlands. This means that data from the careful monitoring of the Rhine in Germany are used by the Dutch authorities to assess flood hazard and provide at least two days lead time before a flood arrives. Bas Overmars points out that the Dutch and German flood models are coupled. This means that Dutch authorities work closely together with their provincial and municipal counterparts in neighbouring North-Rhine Westfalia. So, flood modelling and evacuation planning occurs on the basis of a flood-system management rather than on the basis of political boundaries.

In countries such as the Philippines lead times are much shorter. This is because catchments that are the source of flooding tend to lie much closer to the areas that become flooded. In the Binahaan watershed warning times are only between about 3 and 10 h, depending on the location of settlements along the river. The warning time could be increased if Earth observation was used to monitor actual rainfall, instead of waiting for river gauges to show rising waters. For example, geostationary weather satellites, such as GOES or Meteosat Second Generation (MSG), estimate rainfall in near-real time in other parts of the world. These also provide data suitable for flash-flood detection. Another instrument specifically designed for estimating rainfall in tropical areas is the Tropical Rainfall Measuring Mission (TRMM). Since 1998 this satellite has been used to measure tropical precipitation and Neussner's group is looking into using its data as well. Thus Earth observation is also well suited to providing critical precipitation input data for modelling the flood potential of a given catchment.

12.2.6   Coordination and communication as part of early warning systems

As long as Earth observation data and flood models are only available to the organization in charge of flood monitoring, complete disaster avoidance or mitigation is not possible. Instead, the warning information needs to be shared amongst all organizations involved and, ideally, integrated with other relevant data, such as that on evacuation routes or sites where hazardous materials are used. In the Netherlands, the freely-available National Risk Atlas (`http://www.risicokaart.nl`, see also Figure 12.9) provides such risk-related base information. For real-time emergency situations, the Flood Information and Warning System (FLIWAS, `http://www.fliwas.eu/`) allows real-time integration of flood-related measurements and risk calculation and provides a decision-support tool for evacuation planning. People can also use an internet-based Evacuation Calculator to find out where they should evacuate to, and how long it will take them. This has been integrated into the Dutch Flood Management System. With the many uncertainties crisis managers face in an emergency situation in mind, diverse spatial information has been integrated into an evacuation scenario modeler (Evacu-Aid, Figure 12.10) that allows planning of the optimal evacuation strategy.

The Philippines Flood Early Warning System (FEWS) does not have such a level of automation. When precipitation and river-gauge data indicate an impending flood, the operations centre refers to an established flood-warning plan that has three levels, which are also known to the local population (Figure 12.12). Level 1 (alert) means there
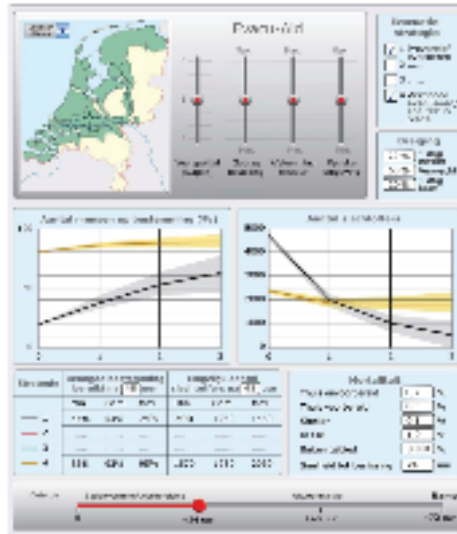
Figure 12.11
Evacu-Aid tool to assist crisis managers in the Netherlands when making evacuation decisions.
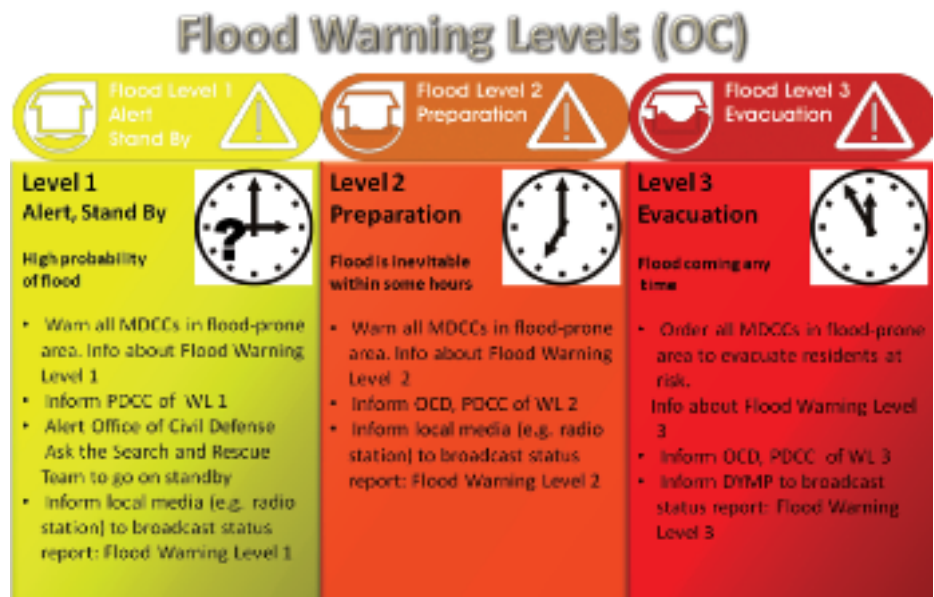
Figure 12.12
Simple three-level flood early-warning and evacuation plans that are part of the Binahaan flood early-warning system in the Philippines [81].

is a high probability of flooding, Level 2 (preparation) means that a flood is inevitable within several hours, while Level 3 (evacuation) indicates that flood waters can arrive at any moment. The operations centre then alerts disaster coordinating councils in the towns and villages. The actual warning of households is still frequently done via a megaphone or bells, a simple but effective system, and is the final stage in a process that relies heavily on RS and other GI Science tools. Both Neussner and Overmars point out that something as low-tech as drills are still critical. It should be ensured that everyone really understands how to behave in a situation of crisis but also to increase and retain relevant skills and knowledge amongst the cooperating organizations.

12.2.7 Future developments in disaster risk management

GI Science data and tools are already indispensable in disaster-risk management—and not just for flooding. Many people use them, often without any detailed understanding of the underlying science and technology. Olaf Neussner and Bas Overmars illustrated two rather different approaches to flood-risk management, although both rest heavily on geo-tools and both involve professional and lay users.

We can expect existing setups to improve, in different aspects and for different reasons. In the Netherlands even more detailed LIDAR-based elevation model called AHN-2 is nearly completed at the time of writing this book (summer 2012). This will lead to elevation data of 5 cm elevation accuracy. With computers becoming faster and models becoming better, flood models will become more accurate and more detailed. Increased availability of broadband internet access will enable an effective exchange of data and information, as well as real-time collaboration between organizations responding to disasters. It also means that the general public will have increasing access to diverse hazard-related information. This will have positive effects, such as an increase in general awareness of existing risks or developing hazardous situations. People will become more familiar with graphic representations of geodata. On the other hand, it does mean that crisis-management organizations will no longer be exclusive users of GI Science and that they will need to adapt their early warning and evacuation strategies to take into account situations where the affected population may be exposed to potentially conflicting information from different sources.

Countries such as the Philippines will benefit from better, more easily available and affordable image data, but also from free and open source software that frees up budgets for other components of the system. The increasing use of sophisticated EO data and flooding models shows the important role that modern geodata and tools can play. Until a few years ago, this was unthinkable in more peripheral settings such as the Binahaan area. Increasing access to GI Science products distributed via digital media (e.g. internet or mobile phone) also means that the number of lay users of GI Science in such economically less-developed countries is growing rapidly.

12.3 Monitoring coastal vegetation

12.3.1 Introduction to monitoring coastal vegetation

Sea-level rise combined with storm surges are potential hazards for communities in low-lying areas of countries such as the Netherlands or Thailand. From a historical (geological) perspective, the current rise in sea levels appears to be decelerating. Sea levels were -100 to -120 m during the last Glacial Maximum at 18,000–20,000 years BP, -15 m at 8000 years BP and -10 at 6000 year BP [62]. In other words, the rate of sea level rise has decelerated from about 0.60–1.0 m/century to about 0.10–0.25 m/century (or 1–2.5 mm/year) at present.
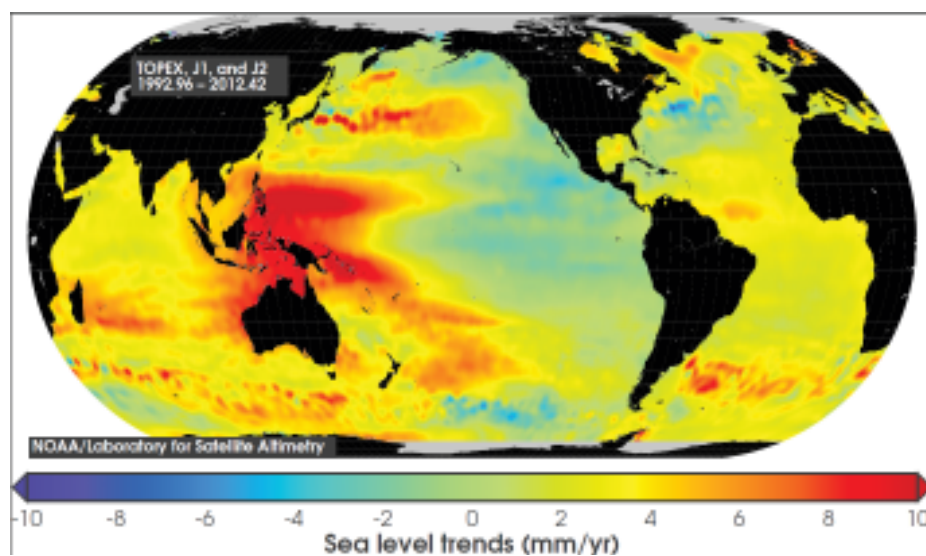
Figure 12.13
Sea level rise based on measurements from satellite radar altimeters. The local trends were estimated using data from TOPEX/Poseidon, Jason-1, and Jason-2, which have monitored the same ground track since 1992. Source: [83].

The IPCC estimates that the global average sea level may rise between 0.18 and 0.59 m in the next century [88]. Sea level changes are not equally distributed in space as can be seen in Figure 12.13. The effect of sea level rise may be accelerated in some regions while in other areas they may not. Acceleration occurs in the western part of the Netherlands and the southeastern part of the UK, as a result of gradual land subsidence as other land is lifted higher due to isostatic rebound after the disappearance of the glacial ice sheet from the last glacial period. This sinking of the land surface is further compounded in the Netherlands by:

- the annual extraction of natural gas (varying between 68 billion m^3 and 86 billion m^3 annually in 1990–2010) and petroleum (in 2007, 2.5 million m^3), which causes local and regional land subsidence of several decimetres [23] in areas above the gas fields;

- land subsidence of peat areas in the western part of the country, owing to drainage and subsequent oxidation of organic matter, where the extent of the subsidence is determined by drainage depth and the thickness of the peat layer.

As a result, some local regions may experience changes in sea level at a higher rate than the 0.18–0.59 m/century predicted by the IPCC, although other regions in northern Europe, such as Sweden and Denmark, are still rising after the unloading of the ice sheet; in these areas sea level is falling.

Coastal flooding occurs periodically along the European coastline. It was a particularly severe combination of high spring tides and a wind storm that resulted in the North Sea flood of 1953 (see Section 12.2). Fifty years later, an earthquake in the Indian Ocean triggered a tsunami with waves up to 30 m high along the coasts of most countries bordering the Indian Ocean. The height of a tsunami's waves is determined by a number of factors, including the size of the earthquake, the amount of ocean sediment displaced by the earthquake, and the length of the downslope along which the sediment travels. More than 225,000 people were killed in fourteen countries [110]. The damage and losses from the tsunami were substantial and coastal and rural communities suffered disproportionately high impacts, requiring central government assistance for their recovery.

Figure 12.14
Two Quickbird images of the harbour area at Banda Aceh, Indonesia, showing the damage of the tsunami of December 26, 2004. (a) 23 June 2004, (b) 28 December 2004.

This disaster prompted a series of studies to examine whether soft defences provided by coastal mangroves are able to protect shoreline areas. Danielsen et al. [25] showed that mangroves give significant protection to some shorelines, where densities of more than 30 trees per 100 m^2 may reduce the maximum flow of a tsunami by more than 90%. In other words, sufficient width of intact mangroves can provide protection against tsunamis, while a degraded or widely spaced belt of mangroves offers little or no protection. Modelling shows that not only tree density is important: wave height is also critical. For example, the reduction effect of dense mangrove forest is decreased when waves are higher than 3 m. Indeed, waves greater than 6 m will destroy a mangrove forest. Then, there is no protective effect. The magnitude of energy absorption of tsunami waves by mangroves also depends on the stem and root diameters of trees, shore slope, bathymetry and tidal stage. Nevertheless, in this context Kerr et al. [56] concluded that "...the apparent association of vegetation area on mortality is in fact due to a tendency for more vegetation to occur at higher elevations and, not surprisingly, to the greater potential areal extent of vegetation given more available area fronting a hamlet. In other words, given hamlets of equal elevation and distance from the sea, differences in vegetation area did not mitigate human mortality caused by the tsunami...We see a genuine danger in overstating the protective capacity of vegetation, because it may lead to a false sense of security and eventually, when the next wave comes, to a lack of trust in science."

In response to such natural disasters resulting from tide and wave damage, the Netherlands and Thailand have developed different strategies for the monitoring, protection and strengthening of coastal defences. These defences (shown in Figure 12.15) include *soft*, or natural, barriers such stabilized sand dunes or mangroves and *hard* constructions built by engineers. Examples of the latter are dikes or tidal surge barriers such as those discussed in Section 12.2. Not only are soft barriers important for protection against storms and waves, coastal wetlands and dunes also function as remnants of natural ecosystems. Salt marshes are often valued as sinks for organic material, nutrients and heavy metals, since sedimentation of particulate matter takes place there; and for their important role in nutrient cycling, which contributes to water quality.

Salt marshes and coastal dunes also provide habitats for wildlife and are attractive natural environments for tourists.
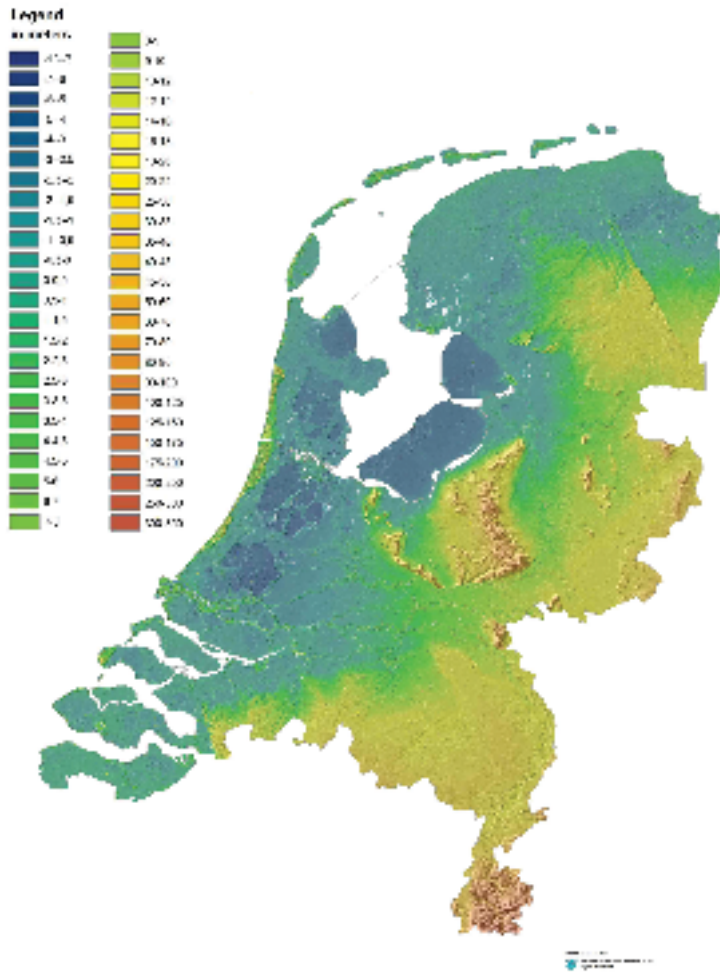

Figure 12.15
Hard and soft barriers along the coast. Westkapelle, small town in the province of Zeeland in the south west of the Netherlands, is protected by dikes and dunes. Source: Rijkswaterstaat, `https://beeldbank.rws.nl`

To understand the close relations between coastal defence and nature conservation in the Netherlands it is important to realize that more than one-third of the country lies below sea level (see Figure 12.16). The Dutch coast consists for 80% out of sand dunes, locally in combination with extensive salt marshes and tidal flats, which, like mangroves elsewhere in the world, are part of the land's coastal defence system. Apart from an obvious role in flood protection, these coastal ecosystems play an important role in conserving national and European biodiversity. For example, more than half of current Dutch flora can be found in coastal dune areas, with approximately 10% occurring exclusively in these areas. For this reason, large parts of the Dutch dunes and salt marshes are nature reserves (see Figure 12.17).

To understand the close relations between coastal defence and nature conservation in the Netherlands it is important to realize that more than one-third of the country lies below sea level. The Dutch coast consists for 80% out of sand dunes, locally in combination with extensive salt marshes and tidal flats, which, like mangroves elsewhere in the world, are part of the land's coastal defence system. Apart from an obvious role in flood protection, these coastal ecosystems play an important role in conserving national and European biodiversity. For example, more than half of current Dutch flora can be found in coastal dune areas, with approximately 10% occurring exclusively in these areas. For this reason, large parts of the Dutch dunes and salt marshes are nature reserves (see Figure 12.17).

A similar situation is found along the country's main river systems. Melting snow or heavy rains in the upper catchment previously led to flooding in the Netherlands. In fact, 65% of the country would suffer from regular flooding if there were no dikes or dunes. The oldest dikes in the Netherlands were built in the 10th century and organized dune management dates back to the early 13th century. It is fair to say that the present shape of the Netherlands is the result of human activity. Another similarity with the coastal fore-dunes can be seen in the fact that many wetlands outside

Figure 12.16
Elevation map of the
Netherlands. Source:
Rijkswaterstaat making use
of the database "Actueel
Hoogtebestand Nederland
(AHN)"

the river dikes (in Dutch, uiterwaarden) have a nature conservation function and are
subject to EU directives such as the European Water Framework Directive (EWFD) and
Natura 2000. Nature development in these riverine wetlands is inextricably connected
to flood-risk management.

Major cities such as Amsterdam, Rotterdam and The Hague, which together contain
approximately 50% of the Dutch population, are located near the coast. The coastal
region also contains major economic and industrial centres such as Rotterdam Euro-
port and Schiphol International Airport and is the source of approximately 50% of the
Dutch Gross National Product [17].

One of the prime tasks of "Rijkswaterstaat" (the Directorate-General for Public Works
and Water Management), a department of the Ministry of Transport, Public Works
and Water Management is to provide and organize protection against floods, so all
coastal fore-dunes and salt marshes fall under its direct responsibility. In order to meet
both coastal defence and nature conservation targets, Rijkswaterstaat has been using
RS and GISs to map and monitor dunes and other coastal ecosystems since the early
1970s. For example, within the framework of the VEGWAD programme, every five
years the vegetation of all salt marshes in the Dutch Wadden Sea region are mapped at

Figure 12.17
The Dutch ecological
network: a planning for areas
with a nature conservation
function. Source: [45].

a scale of 1:1000 and stored in a GIS for monitoring purposes. Initially the maps were based on large-scale aerial photographs, but in recent years more advanced techniques such as Laser altimetry and other airborne scanners are being used (see Section 12.2).

Since Rijkswaterstaat is one of the first and largest users of RS techniques, for many years it has played a leading role in the RS and GIS community in the Netherlands. It has initiated and (co-)funded numerous EO research and operationalization projects, covering the whole spectrum of available EO techniques with possible (future) applications directly related to the primary tasks its home ministry.

As already stated, dunes and other coastal ecosystems have multiple functions. This has as a consequence that mapping and monitoring projects are always executed in close cooperation with other stakeholders, among them Government, NGOs, knowledge centres and national and local authorities.

Rijkswaterstaat incurs significant annual expenditure to map and monitor vegetation across a number of landscapes and ecosystems within the Netherlands. A land unit approach, based on visual interpretation of stereoscopic aerial photographs, and supported by field observations, has been developed by the Survey Department (Meetkundige Dienst). Though this approach is effective, Rijkswaterstaat wished to investigate fur-

ther whether hyperspectral digital RS in combination with laser altimetry would increase the productivity and objectivity of the mapping procedure. ITC and Rijkswaterstaat developed a project that delivered a marked improvement in mapping.

Initially, 16 salt-marsh vegetation types in the Netherlands were identified based on existing knowledge within Rijkswaterstaat, each type having a characteristic spectral signature (see Figure 12.18, [103]). In addition, an understanding was gained about those parts of the electromagnetic spectrum that offer the greatest information content for discriminating between and identifying vegetation types.
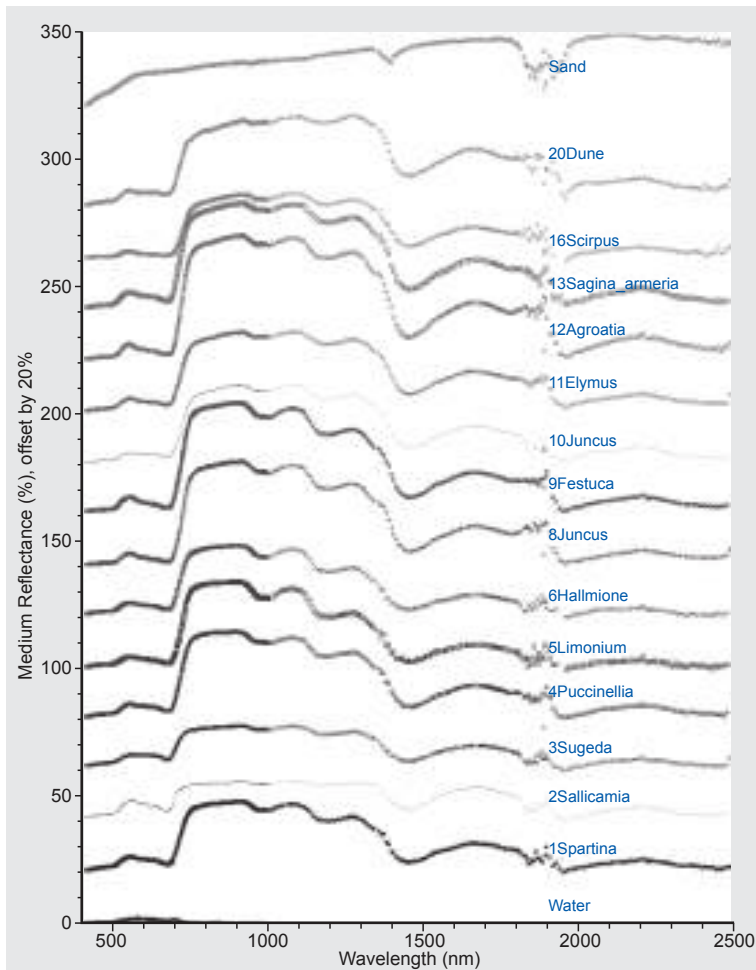
Figure 12.18
Reflectance profiles of
salt-marsh vegetation types
in the Netherlands [103].

From knowledge gained from ecologists working on coastal vegetation on the island of Schiermonnikoog, five input map layers were identified as being important for explaining the distribution of coastal vegetation. The first was a spectral-angle-mapper (SAM) supervised classification of hyperspectral (HYMAP) images. The other map layers, which were derived from a digital elevation model, represented elevation, slope gradient, aspect and topographic position. These layers were input in a raster-based GIS and then geometrically co-registered to a regular 3.5 m grid. From knowledge of vegetation distributions, the relationships between the vegetation units and the five data layers were quantified and used as rules for a rule-based expert system. The thematic layers accessed from the GIS provided data for the expert system to infer

the most-likely vegetation unit occurring in any given grid cell. The vegetation map output from the expert system compared favourably with a conventional landscape-guided map generated from aerial photograph interpretation [104].

The main conclusions to be drawn from the testing of the technology developed to produce maps of the Dutch salt marshes were:

- the accuracy of vegetation maps produced using conventional aerial-photograph interpretation was 40%;

- the spectral-angle-mapper (SAM) algorithm used in combination with hyperspectral imagery was able to classify vegetation with an accuracy of 40%;

- ecological knowledge is available from field plots, as well as expert ecologists;

- ecological knowledge linking environmental variables to vegetation types could be captured in a set of expert rules;

- the addition of terrain variables to the hyperspectral imagery raised the accuracy of the map generated by the expert system to 62%;

- the overlaying of polygons from a 1997 vegetation map on the results of the expert system showed that, during the interpreting of aerial photographs, much of the ancillary information (elevation, slope, terrain position) was used when drawing lines to denote boundaries of vegetation units.

12.3.2 Solutions and future developments in coastal vegetation monitoring

The vegetation of coastal areas provides a range of environmental benefits. As a soft defence, it can be combined with hard defences to protect human infrastructure from storm events. GI Science provides essential tools for mapping soft defences, monitoring their condition and developing models of their future effectiveness under different land-cover scenarios and projected storm-surge impacts. Ongoing and future initiatives focus on improving mapping techniques and implementation of these in coastal defence systems as well as in nature conservation.

The ITC professional Masters students in the NRM programme of 2010 focused in their project work on mapping the habitat quality of the hen harrier. The hen harrier is a bird species whose decline and conservation status in Europe is of great concern. Because the hen harrier depends for foraging and breeding on a mosaic of open and closed vegetation, the ITC students mapped the vegetation structure on Schiermonnikoog (one of the Wadden islands in the north of the country). They collected vegetation structure samples in the field and classified digital false colour aerial photographs. One of these students, Sylvia Monica Kalemera, compared different methods of classifying vegetation structure based on RS techniques. See also Figure 12.19. Further improvement in classification accuracy will allow better predictions and more efficient coastal management to ensure both safety for people and protection of the hen harrier habitat.

ITC graduate Giles Jay Williams [128] estimated chlorophyll content in mangroves in East Kalimantan, Indonesia, based on hyperspectral techniques (Figure 12.20). Mangroves in Indonesia have suffered from logging and nutrient enrichment due to unsustainable shrimp farming and agriculture. Therefore, mangrove monitoring and measures to improve the state of the mangroves are necessary. In view of monitoring mangroves, chlorophyll content has a strong correlation with nitrogen and can therefore be expected to be a good indicator of mangrove health.
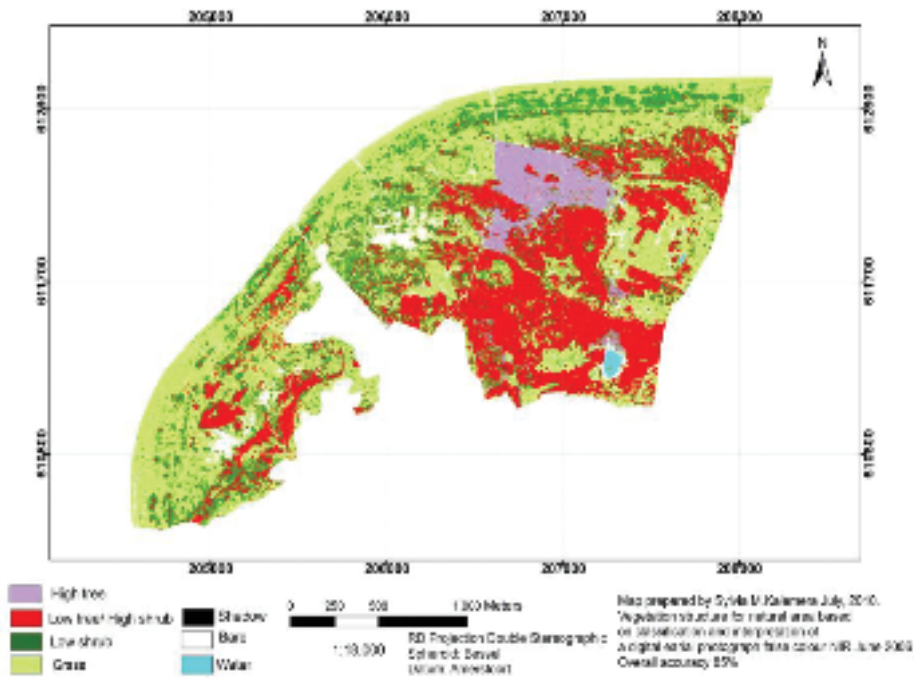
Figure 12.19
Vegetation structure map of a part of Schiermonnikoog, the Netherlands, in support of mapping the hen harrier habitat quality.
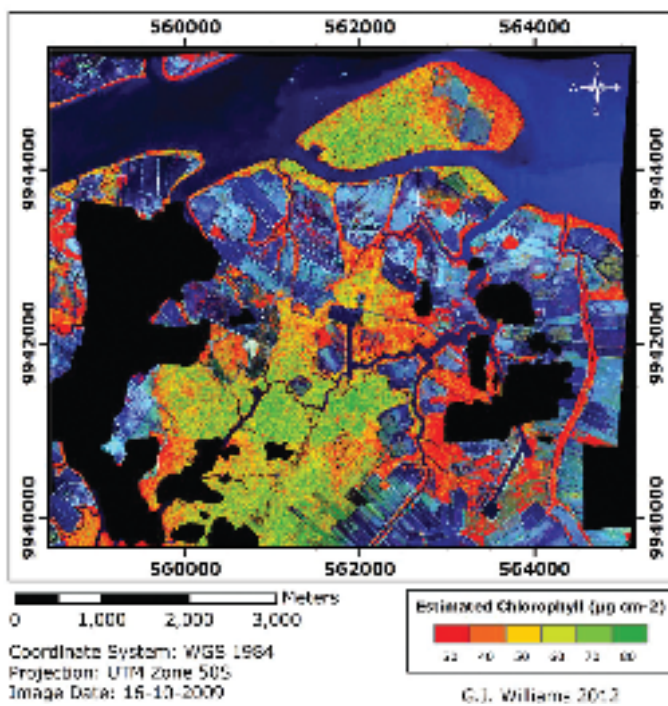
Figure 12.20
Chlorophyll mapping in a degraded mangrove area in East Kalimantan Indonesia. Source: [128].

Based on this study it was concluded that the created maps do not reflect accurately chlorophyll at specific locations. Several methodological problems were identified that

455

need further attention. Nevertheless, such maps are indicative for determining spatial variations throughout the study area. Chlorophyll concentration was correlated with shrimp pond proximity and position relative to the coast.

Ongoing sea level rise and degradation of natural vegetation along coasts asks for well-targeted protection and conservation measures. That will only be possible with good-quality baseline information. Both recent studies show that further development of innovative EO and spatial analysis techniques is the way forward to monitor coastal vegetation with a higher level of accuracy.

12.4 Nature conservation

12.4.1 Introduction in nature conservation

Nature conservation aims to conserve nature areas and the ecosystems they contain. If not properly protected, nature areas could be converted into agricultural land or subjected to urban sprawl. Or they could be utilized to an extent that structurally damages the functioning of the ecosystem, for example through deforestation or over-grazing. Sometimes, "new" nature is created when agricultural areas are abandoned, when land is reclaimed from the sea, or when reforestation projects are carried out in degraded areas. Then, these areas also need proper management and protection. Overall, in all nature areas, nature conservation aims at ensuring that biodiversity in these areas is maintained, that natural processes in the system continue and that ecosystem benefits are retained.

Two situations are often found in nature areas. In the first situation, nature areas cover large expanses such as African game parks. In the second situation, nature areas are embedded in a land-use matrix of areas that are intensively or extensively used by humans. In the latter case, agricultural areas often surround the nature areas, although urban sprawl sometimes surrounds a park (e.g. Nairobi National Park). In both situations, to effectively conserve these areas, spatial information is essential for making the right management and conservation decisions. Spatial information can help in monitoring the status of the system, prioritizing areas requiring the most attention or investigating the connectivity and remoteness of isolated nature patches in the landscape matrix.

Collecting relevant spatial information for nature conservation is, however, not an easy task. Firstly, nature areas can cover large expanses of land or be embedded among other types of land use. In either case information covering large areas is required. Earth observation from aircraft and satellites has revolutionized the way data from large areas are collected in a consistent and repeatable manner. Secondly, changes in nature areas are often slow, making those changes difficult to observe instantaneously. Collection of long (in the order of decades) series of data is then necessary to quantify the rate of change in a natural system. Historical archives containing aerial photos and old satellite images provide a valuable resource for assessing the state of natural systems in the past. Finally, collecting field information based on point observations is often only useful when the point's exact geographical position is known—to be able to relate it later on to other spatial data sets. In nature areas, however, landmarks are not always readily available, making it difficult to pinpoint one's position. With the advance of small hand-held devices that can receive GPS signals, this task has been considerably simplified and accuracy improved. In short, geographical information is essential for nature conservation and collecting this kind of information has become easier and its quality higher with the advance of EO sensors and GPS devices.

12.4.2 Users and user requirements related to nature conservation

Nature conservation usually involves various stakeholders. To facilitate discussion between all stakeholders, to quantify the impacts on and from nature areas, and to prioritize areas and actions by the different stakeholders, nowadays it is essential to have access to spatial data from the nature areas and their surroundings. As they provide the means for processing that spatial data, GI Science tools have acquired a pivotal role in optimizing nature conservation.

Typical stakeholders in the field of nature conservation are non-governmental organizations (i.e. NGOs such as WWF and Conservation International). Other important players are national and lower-order governmental bodies (e.g. a state forestry de-

partment) and international bodies that originate from supra-national organizations such as the UN (e.g. the International Union for Conservation of Nature IUCN). These kinds of parties all have conservation of nature as a common goal.

There are other stakeholders that make claims on nature areas but they do not have nature conservation as one of their primary objectives. Indigenous people, for example, often claim land-use rights or ownership of areas that have been assigned a nature conservation status. Additionally, within a national government different ministries can claim responsibility for, and therefore authority over, the same piece of nature, e.g. a ministry of agriculture and a ministry of forestry. Next to various parties that make direct claims on using or managing nature areas, there are many parties that live next to these areas that are directly or indirectly affected by the natural processes occurring in these areas. For example, cattle herders living next to game parks sometimes experience loss of livestock by predators from the neighbouring game park, representing a negative effect of nature areas on surrounding communities. Positive examples of nature areas exist as well, for example where villagers are protected from landslides by forested slopes that retain and regulate the runoff of rainwater in a catchment area.

Surrounding communities often have a negative impact on nature areas, caused, for example, by poaching or illegal collection of fuel wood. But positive benefits also exist, for example where communities earn revenues as tourist guides and in return actively help in protecting the area. Finally, often research organizations also have an interest in nature areas: they often try to gain access to data from these areas and acquire permission to perform experiments and gather observations from within these areas.

For stakeholders with nature conservation as their primary objective, GI Science provides important tools for conservation management. In their discussions with other stakeholders, it often serves as a useful tool for communication and for making arrangements, e.g. to negotiate access rights to specific areas. And for the research community, spatial information often helps in finding relationships between natural processes that occur at the same locations.

Applications of GI Science in nature conservation are extremely diverse. Therefore it is impossible to describe all of them. This Section illustrates three examples of use and users of GIS and RS applications. Each of them deals with one of the major stakeholders in nature conservation: an NGO (Subsection 12.4.3), a government organization (Subsection 12.4.4) and a research organization (Subsection 12.4.5).

12.4.3   Use of spatial information by an NGO: Natuurmonumenten's vegetation-structure map of Witte Veen.

Natuurmonumenten is an NGO with 880,000 Dutch members. They manage over 100,000 ha of nature areas in the Netherlands. One such nature area is the Witte Veen, 10 km south of Enschede, which links up with a nature area in Germany. Hans Gronert, employed by Natuurmonumenten, explains: "Together with German colleagues we are trying to connect this area to the nearby nature areas of Aamsveen and Haaksbergerveen, as well as other neighbouring nature areas. Our management approaches are grazing with Scottish Highland cattle to keep the area open, removal of topsoil to maintain the nutrient-poor environment and the promotion of frog pools in agriculture areas as stepping stones for amphibians. To be able to set up our management plans and to monitor their effect we need vegetation-structure maps. A few examples of rare species that can be found in the Witte Veen area are the European tree frog, common cotton grass and blue gentian."

Hans Gronert, a forester employed by Natuurmonumenten, is involved in nature management of the nature area of Witte Veen. His main task is to optimize and maintain the biodiversity of this area.

The Province of Overijssel needed to create new nature areas to establish the National Ecological Network (EHS), which has been set up to connect existing nature areas in the Netherlands. A large part of existing and future nature areas have been assigned as Natura 2000 areas. Natura 2000, the centre piece of EU nature & biodiversity policy, is an EU-wide network of nature protection areas established under the 1992 *Habitats Directive*. The aim of the network is to assure the long-term survival of Europe's most-threatened and valuable species and habitats. It comprises Special Areas of Conservation (SAC) designated by Member States under the Habitats Directive and also incorporates Special Protection Areas (SPAs), which they designate under the 1979 *Birds Directive*. Natura 2000 is not a system of strict nature reserves where all human activities are excluded. Although the network will certainly include nature reserves, most of the land is likely to continue to be privately owned and the emphasis will be on ensuring that future management is sustainable, both ecologically and economically. The establishment of such networks of protected areas is an obligation that the Netherlands need to fulfil under the UN Convention on Biological Diversity. Natuurmonumenten, the state forestry department, the provincial government dealing with nature areas and other conservation organisations work together to achieve this.

One of the requirements for managing nature areas is baseline information such as insight in the spatial distribution of different vegetation types. Vegetation structure is an important habitat characteristic for many species and in support of biodiversity it is important to maintain diversity in vegetation structure. The vegetation structure map of the Witte Veen is shown in Figure 12.21. This map was produced by visual interpretation of an aerial photograph mosaic downloaded from Google Earth.

12.4.4 Use of spatial information by a government organization: the forest-cover map of Rwanda

The National Action Plan for Forests 2006–2008 in Rwanda was aimed at research and operationalization of the use of RS in forest inventories. More specifically the plan prescribed mapping forest resources in Rwanda at 1:25,000 scale. This included:

- making an inventory and estimate the location, area (of at least 0.5 ha in size), floristic composition, age, type (natural or not; public or private), soil type, density and health status of forests/woodlands using existing maps, aerial photography, satellite imagery and field data collection;

- monitor changes in the occupation of forest land over time;

- inventory, organize, standardize and centralize national geographic and other available databases on forests in Rwanda to allow interested institutions and

Natura 2000

Figure 12.21
Delineation of mapping units (yellow) based on interpretation of an image obtained from Google Earth (left) and vegetation-structure map (right) after data collection in the field using mobile GIS.
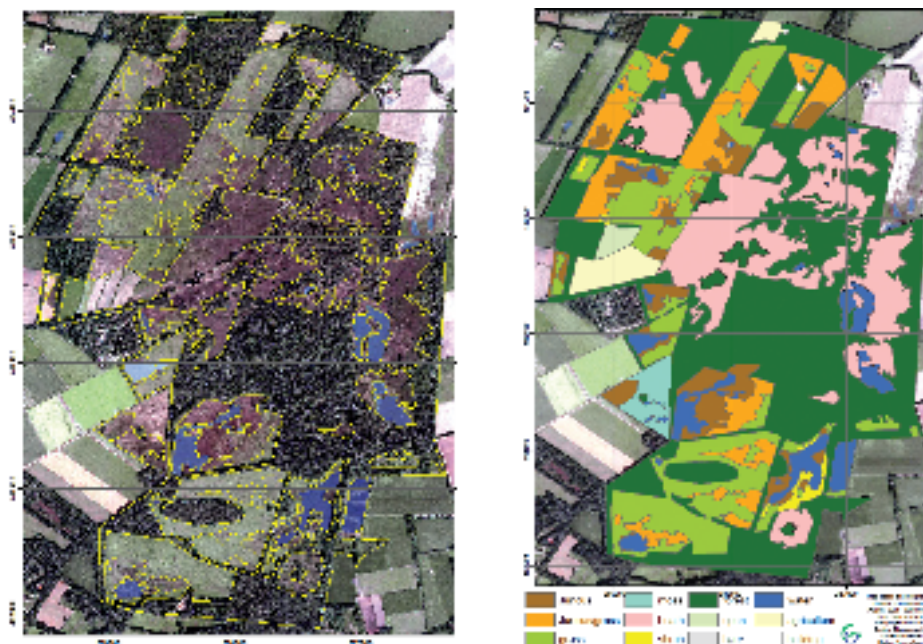
decision-makers to easily access and update information that is critical for decision-making processes;

- develop a national GIS/RS-based information system for forests in Rwanda;

- develop capacity building in the application of RS and GIS tools and methods for forest inventory and mapping.

Claudien Habimana (right), Director of the Forest Unit, Ministry of Lands, Environment, Forestry, Water and Mines (Minitère), Rwanda, is listening to an explanation being given by a local farmer (left).

The satellite data used to map Rwanda's forests were ASTER, SPOT and TM images, the most recent (at the time of the described project) and cloud-free images being selected (see Figure 12.22).

From discussions with the Ministry officers, criteria for distinguishing forest types were drawn up. Forests in the humid region of the country were defined as areas larger than 0.5 ha with a tree cover greater than 20% and trees higher than 7 m. In the dry region (Akagera), areas in which trees were higher than 5 m were considered as forest. Coppices and young forest plantation, the latter comprising trees of less
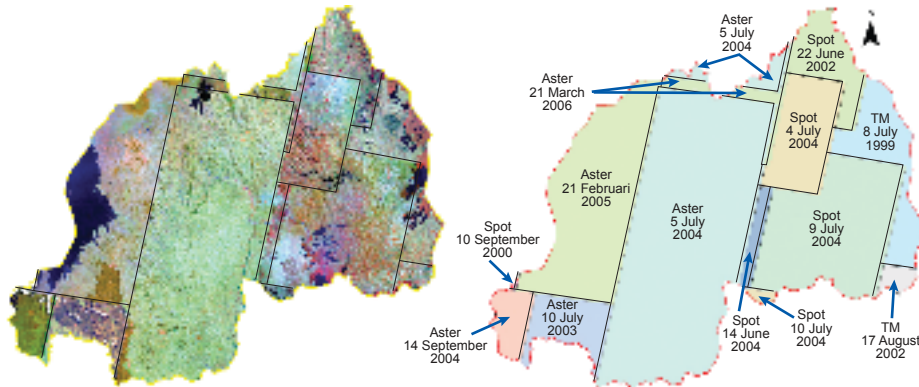
Figure 12.22
Coverage of most recent
satellite images (at the time
of the project) used for the
forest classification in
Rwanda.

than 7 m and covering 10–40% of the area, or trees higher than 7 m and covering 10–
20%, were only found in the humid region. Bush was only found in the larger humid
natural forests. In the case of bush, tree cover was defined to be less than 20% and
in practice mostly less than 10%, with shrub covering more than 20% and high herbs
and grasses. Bamboo and bush ridge forests were only found in the Parc National des
Volcans. In summary, the following forest types were distinguished:

1. humid natural forest;

2. dry natural forest;

3. eucalyptus plantation forest;

4. pine plantation forest;

5. coppices or young forest plantation;

6. bush;

7. bamboo forest;

8. bush ridge forest.

Because of the substantial differences in terrain elevation, the images needed to be
corrected with a DEM. Geocoding was done from topographical maps in accordance
with the Rwandan coordinate system. The images were each classified separately with
ERDAS IMAGINE into several classes and then re-coded into two classes: *forest* and
non-forest. The large forest areas were selected from this classification and classified
separately into the forest types listed above. Field data collection was carried out
using hand-held GIS and GPS devices, allowing the location to be seen in the field
in relation to the polygons, with the satellite image in the background. These data,
collected over 4 missions, were classified into the forest-cover classes.

Based on the field observations and a first classification of images, the definition of
classes was adjusted and a completely new classification of all images was made. New
spectral signatures were selected and several classifications were made until the forest
was sufficiently accurate classified. The occurrence of each forest type, expressed in
km^2, is shown in Table 12.1.

Table 12.1
The combination of natural forest, bamboo forest, bush ridge forest and forest plantations can be considered collectively as "forest". Hence, the proportion of forest covering the total area is 6.8%; bush cannot be considered as forest.

Forest type	No. of polygons	Area (km^2)	Fraction (%)
Non-forest	699	23,252	91.9
Bush	5584	343	1.4
Bamboo forest	8	44	0.2
Bush ridge forest	5	30	0.1
Dry natural forest	664	37	0.1
Humid natural forest	1430	798	3.2
Eucalyptus forest plantation	1164	306	1.2
Pine forest plantation	663	110	0.4
Young forest plantation or coppices	22768	392	1.5
Total	43,426	25,312	100.0
All forest	37,143	1717	6.8

12.4.5 Use of spatial information by a research organization: monitoring tree-line movement resulting from climate change and societal pressure

Users are interested to see what the state of their forests is for a variety of reasons, biodiversity management and carbon sequestration being only two. Carbon sequestration has become an important issue for forest managers because forests contain the bulk of the carbon of global terrestrial ecosystems and keeping the carbon in forests is an important measure for reducing greenhouse gas emissions. Increases in atmospheric CO_2 concentrations, and global warming of up to 2 °C, however, seem unavoidable, and these may also affect forest ecosystems. CO_2 fertilization (plants need CO_2 for photosynthesis) has a positive effect on plant growth. At higher latitudes and in mountain regions, where temperature limits tree growth, forests may also expand as a result of global warming. This would result in tree lines moving to higher altitudes. To monitor tree lines, RS products are ideal sources of information. Progression of tree lines and the rate at which they move can be calculated in combination with accurate DEMs and GIS systems.

The pastures around Osogovo Mountains, on the border between Macedonia and Bulgaria, provide a good example of tree-line movement. Bulgarian researcher and forest manager Dr Tzvetan Zlatanov is keen to know whether forests in this region will expand to higher latitudes as a result of global warming. This information is needed for at least two reasons. Firstly, for reporting in connection with the Kyoto protocol—the country has to account for changes in carbon stocks in their terrestrial biomass. Secondly, forest expansion in the mountains will occur at the expense of unique types of alpine vegetation and will, thus, have consequences for biodiversity. This needs monitoring too.

Dr Tzvetan Zlatanov, Forest Research Institute, Sofia, Bulgaria

A complicating issue is the societal change that is taking place in Bulgaria at the same time as climate change. Before the country became part of the Eastern Bloc (also called Communistic Bloc) countries in the 1950s, all pastures around the Osogovo Mountains belonged to private owners. Then in the 1950s, all land became state owned and most of the previous owners moved to the cities to take jobs in the country's industrial sector. When, after 1989 the democracy was reinstalled in Bulgaria, land was given back to its previous owners, but many of them had lost the know-how needed to manage their pastures properly. As a result, many of the pastures became abandoned land. Proper pasture management includes frequent mowing and summer grazing, which have adverse effects on tree establishment. When lands are no longer used, tree settlement is no longer hampered, allowing forest expansion (see Figure 12.23).

In order to monitor the movement of the tree line to higher altitudes in the Osogovo Mountains and to relate it to climate change, Dr Zlatanov has to correct for the effects

Figure 12.23
Garlyansko Zhdrelo Gorge in the Osogovo Mountains. Newly recruited trees above the tree line stand out clearly.

of land abandonment. On the other side of the Bulgarian border, in Macedonia, land abandonment has not taken place, because there the land always had remained private property. This area of the Osogovo Mountains is therefore ideal for a baseline study.

Calculating the rate at which the tree line moves requires high-quality DEMs and aerial photography. Historical photos of mountain slopes can also be very useful, as they give clear indications of where the tree line was previously positioned on the mountains. Together with recognizable features on slopes, the position of the tree line can be delineated in a GIS system, allowing further analyses. Differences in tree-line movement on either side of the Bulgarian–Macedonian border can then be attributed to differences between the socio-economic development of the two countries. In addition, the rate of change in Macedonia gives a clear indication of the effect of climatic change on tree growth.

12.4.6   Future developments and datasets for nature conservation

GI Science is widely used for nature conservation at various levels of detail, ranging from very specialized local applications to worldwide global investigations. Each level requires different types of spatial data and GI tools in order to achieve the required result. Therefore it is impossible to generalize data requirements of all different users. But, the three examples that have been presented in this section give some idea of typical applications in nature conservation at different scales and for different users.

Within nature conservation, GI Science provides tools for achieving the objectives set; the tools are never the end of a process. The examples also demonstrate that acquisition of spatial data is an important aspect of using GI Science in nature conservation. Data acquisition has been a considerable bottleneck for nature conservation in less-developed countries for many years, as field work for data collection is often expensive. Advances in EO and, over time, a reduction in costs have resulted in the availability of relatively cheap data sets that have been collected in a consistent manner. Nowadays, international efforts are being made to share and disseminate these data at all levels. To complete this section, a summary of some global data sets that are
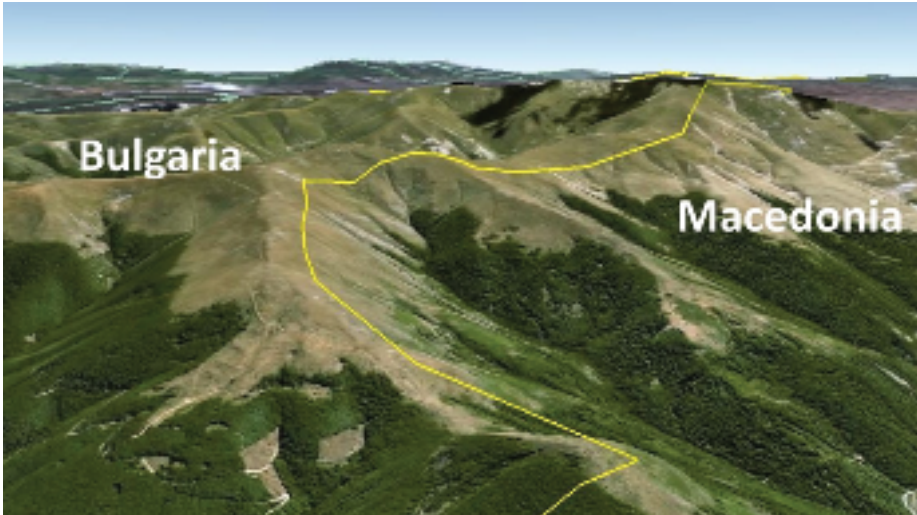
freely available on the Internet seems appropriate:

- Nature-GIS is a network that brings together different stakeholders of protected areas: users and experts in IT and nature conservation; see
  `http://www.gisig.it/nature-gis/`.

- IBAT for business is an innovative tool designed to facilitate access to accurate and up-to-date biodiversity information to support critical business decisions. The tool is the result of a ground-breaking conservation partnership between BirdLife International, Conservation International and the United Nations Environment Programme World Conservation Monitoring Centre; see
  `http://www.ibatforbusiness.org/`.

- The IUCN Red List includes a comprehensive assessment of the conservation status of the world's 6,000+ known species of frogs, toads, salamanders, and caecilians. Also included are key findings of the assessment, as well as individual species accounts, the IUCN Red List threat category, range map, ecology information, and other data for every amphibian species; see
  `http://www.iucnredlist.org/amphibians`.

- The Global Biodiversity Information Facility (GBIF) is an international organization that is working to make the world's biodiversity data accessible everywhere in the world. GBIF and its many partners work to mobilize the data, and to improve search mechanisms, data and meta-data standards, web services, and the other components of an Internet-based information infrastructure for biodiversity; data can be found and downloaded at
  `http://data.gbif.org/welcome.htm`.

- The Food and Agricultural Organization (FAO) of the UN provides a lot of statistics at country level; most data can be downloaded from the FaoStat website, at `http://faostat.fao.org/site/291/default.aspx`.

- Finally, the Earth Explorer by NASA gives a lot of information on several satellite products, some of which are freely downloadable; go to
  `http://earthexplorer.usgs.gov/`.

12.5 Information Exchange for Spatial Planning

Spatial plans have legally binding consequences for governments and their publics. They provide legal certainty and underpin local development. And they must be up to date, because those plans form the legal basis for what administrators, local government, the public and industry are or are not allowed to do in spatial contexts. Some Dutch municipalities started digitizing spatial plans in the early 1990s. During the nineties, government, industry and citizens increasingly recognized the significant potential of the Internet for the exchange of digital spatial plans. At the turn of the 21st century, the national exchange programme for digital spatial plans (DURP, in Dutch: "Digitale Uitwisseling Ruimtelijke Plannen") was born.

DURP

In this section, we highlight the problems that led to the creation of DURP and describe the legal and technical arrangements designed by DURP partners, in consultation with academia, industry and government, to solve those problems. We also highlight a pervasive characteristic of large-scale computerization programmes in government, both in the North and the South: the solution of old problems may bring new problems to the fore.

12.5.1 Introduction to spatial planning

In the Netherlands, the executive branches of municipal governments are responsible for spatial plans. Spatial plans are consulted before issuing a permit for a new building, or even to add a shed in one's garden. Municipal officers prepare these plans after an exhaustive search for all information pertaining to the land in question. All possible land uses are reviewed in the process, and preliminary plans are checked against the rules and frameworks laid down by the provincial and national authorities. The law stipulates that every citizen can have a say, and submit an appeal if necessary, at any stage during the development of a spatial plan.

In the year 2000, the Ministry of Housing, Spatial Planning and Environment, the Ministry of the Interior and Kingdom Relations, the Association of Water Boards, the Association of Dutch Municipalities and the Association of Provincial Authorities became partners in a national programme called DURP (DURP partners). Their aim was to improve the digital exchange of spatial plans. DURP partners framed the problem with (and solution for) spatial plans thus as follows:

DURP partners

"Do you recognize the following situation? Filing cabinets filled with physical plans and related amendments. Changes often take the form of sketches and paper stickers on the original plan. New municipal employees have a particularly hard time searching for plans before they are able to explain to citizens what they are allowed or not allowed to do on their land. Moreover, these plans are often more than 10-years old and must be updated under the new Spatial Planning Act (WRO in Dutch). If this situation is familiar to you, you will have no doubt as to why digital physical plans are needed."

12.5.2 User and user requirements

Users of spatial plans may be citizens, businesses or government bodies at all levels, including the country's 441 municipalities, 12 provinces, the Council of State and its DURP partners. Each of these users may make different demands on the spatial plans they access.

Citizens and businesses want low-cost access to spatial plans, together with convenience and clarity. For companies scouting within a large region for a new location upon which to expand their business, the search can be excruciating. The individual plans that together cover the region may well be all different. And they are often

only available on paper, yet real-estate agents, notaries and public organizations, but also citizens and businesses, prefer to consult spatial plans on the Internet, instead of having to visit the municipal centres during office hours. Indeed, many users wish to consult not only municipal, but also provincial and national, plans via the Internet. In this general context it is also important to note that users expect, if not demand, that spatial plans, zoning regulations and provincial plans are all compatible.

Government officers charged with spatial planning (planners, lawyers and geo-ICT experts) consult with each other across different levels of government about zoning decisions. Often, other government agencies wish to offer their opinions on such matters. All parties involved are in favour of efficient cooperation, preferably via the Internet. So we are rapidly changing from sending piles of paper from the municipalities to provincial and other government organizations to communication via the Internet.

12.5.3 A new law

To tackle these problems, DURP partners agreed to initiate three new measures: new legislation, new technical standards and a spatial planning portal. On 1 July 2008, eight years after the establishment of the DURP partnership, the new Spatial Planning Act came into effect. The guiding principles for the new law were fewer rules, less central control where possible, and an implementation-oriented approach. "These are no empty slogans; they represent an actual simplification of the decision-making process in spatial planning, with due consideration for such important concepts as legal certainty and democracy. The new act ensures a clear division of labour. It distributes responsibilities and powers among municipalities, provinces and the national government in such a way that each tier of government can represent the interests entrusted to it to the best of its ability. [...] The new act aims at achieving the following objectives, inter alia: more efficient decision-making, improved enforcement and simplified legal protection in spatial planning. In short, the act will create an effective, decisive and goal-oriented spatial planning system in the Netherlands." [124]. All municipalities were required to make their plans digitally available at source by 1 July 2009. Due to implementation delays, however, the date was extended to 1 January 2010.

The new legislation in the *Besluit Ruimtelijke Ordening* and in a *Ministeriële Regeling* featured several innovations: spatial plans and the entire spatial planning process have to be digital; digital spatial plans have to be available to everyone; a paper copy must be made for archiving; when in doubt, the digital version of the plan overrides the paper copy; and, the digital version is considered to be the authentic version of the plan.

12.5.4 New technical standards

In 2008 DURP partners finalized a package of standards to enable the digital production and availability of all planning visions, spatial plans, decisions, regulations and orders [35]. The package consists of three standards: a Standard for Comparable Spatial Plans specifying main groups of land uses and area specifications; an Information Model for Spatial Planning for the new instruments and revised procedures; and a Standard for Accessibility to Spatial Instruments and seven practical implementation guides. Under the standards, to make a plan, for example, certain steps need to be followed:

- start with the application of SVBP where it concerns zoning plans and accommodation of plans;

- use the National Triangular Network when creating geometric objects;

- link objects with the IMRO using the practical guidelines;

- use the Standard for Accessibility to certify the provision of the design plan and the established plan;

- make available and publicize the certified plan according to the rules of the Standard for Accessibility.

12.5.5   New spatial planning portal

In part, the new spatial planning portal, which was launched on 1 June 2008, was set up to fulfil requirements that were laid down in the new Spatial Planning Act, but it nevertheless serves a massive, heterogeneous user population. The portal makes digital spatial plans publicly available via web services, covering the entire country for all tiers of government. "With the portal, the government aims at providing spatial plans to citizens, private organizations and government bodies in a transparent way. The site presents the complete and most recent situation at any location in the Netherlands in a reliable and clear way. Consequently, citizens and professionals will be able to integrally query spatial plans" [125].
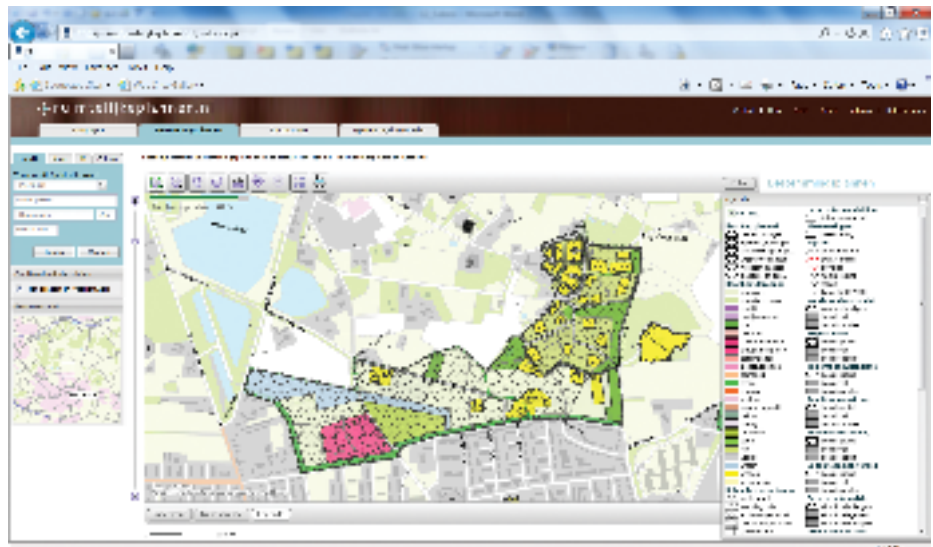
Figure 12.25
The Dutch spatial planning portal. The current spatial plan for a small area at the northern fringe of Enschede is shown. Screenshot was taken on 9 July, 2012 from: [100].

In July 2009, the intended beneficiaries of DURP's activities were asked whether their requirements were being met by the new arrangements. Apparently, these new arrangements had solved some of problems but new ones were created.

Citizens value the ease of querying standardized, comparable spatial plans at all levels of government via the Internet without having to visit various government offices. Municipal planners perceive the increased power accorded to them by the new laws as an improvement. In the past, a municipality drew up local plans and the relevant province approved them. Under the new law, the municipality draws up and approves the plan. In cases of conflict, the province may appeal to the Council of State. Municipalities are content with the reduced decision powers of the province in cases of conflicts with citizens. And provincial planners are comfortable with the power shift to municipalities under the new law. Municipalities decide based on local interests; the province safeguards provincial interests and cannot just interfere when it disagrees with a local plan.

Both the province and the municipality value the efficiency of the appeal process (if the plan needs to be defended in court). In the past, the province had to defend in

court plans made by the municipality that the province was often unfamiliar with. Under the new law, the municipality defends its own planning choices. The power shift towards the municipalities reduces the province's work load and allows it to re-deploy provincial employees (previously tasked with the approval of municipal plans) to other work more relevant to the province (e.g. making local physical plans of provincial interest, such as projects to provide infrastructural elements that span more than one municipality).

Both municipal and provincial planners expect more efficient cooperation under the new law. First, the time needed to develop a plan has been reduced from one year to 22–24 weeks. Second, digital plans enable integration with other spatial data (e.g. plans at the province level). Third, because of its reduced power, the province now has an interest in communicating future provincial plans to municipalities in a timely manner. And in the past, the province could reject a municipal plan that had not been justified in the proper manner, an action the municipality felt was "useless", since it did not have any bearing on the content of plan [36].

Now a citizen can only appeal to the Council of State against an "unfair" municipal decision. This new practice has disadvantages. First, citizens may not trust the municipality to decide "objectively" on their case because municipal employees, who are usually themselves locals, may have a personal interest in certain locations within the municipality. Second, an appeal to the Council of State is not free of cost, in contrast with an appeal to the province, which in the past was cost free. As a result, citizens may be discouraged to exercise their right of appeal. Third, in the past, an appeal to the province had to be processed within six months. Under the new law, the Council of State does not have this obligation and it can take more than a year to process an appeal. In summary, cost and time issues may discourage citizens from appealing municipal decisions they perceive as unfair.

A fourth disadvantage concerns perceived loss of discretionary power. In the era of analogue plans, municipal planners had the freedom to assign any type of land use to a certain area and to apply any system of symbols to the plan. The new standards curtail these freedoms. With the new digital plans, only 23 predefined land-use types are now allowed, while the accompanying symbols are standardized centrally and prescribed by the ministry responsible for planning. Municipal planners may perceive this as limiting their discretionary power to make subtle judgments, based on local knowledge, regarding land use.

Fifth, and finally, provincial planners expect losses of efficiency in cases where they reject a municipality's plan and have to submit an appeal to the Council of State, a process that may take at least a year. The Council of State still works using analogue processes, so provincial planners may need to transpose plans to paper when pursuing an appeal. Under the old arrangements, it was both within the interests and expertise of the province to remove errors before plans achieved a legally binding status. Under the new law, the province has no interest to check for errors. Quality assurance of plans rests entirely with municipal employees, who may lack the technical expertise required for quality assurance, especially in small municipalities.

A similar case of massive government informatization can be seen in the Bhoomi (meaning land) land records project implemented in the state of Karnataka, in India, in 2001. Several pressing problems were solved by the Bhoomi project but new ones emerged. By October 2004, over 22 million farmers had received a digital copy of their land record for the first time. Digital copies of land records can now be obtained on payment of about 30 US cents at decentralized locations (kiosks), where kiosk operators run and maintain the system at a local level; long waiting periods, the need to make several visits and *unofficial payments* to intermediaries are things of the past. For

every transaction, kiosk operators authenticate themselves with bio-logon metrics on Indian-made machines that look a lot like an ATM and are easy to use. The Bhoomi project improved the quality of service to citizens, simplified the administration of land records, achieved financial sustainability and curbed corruption. The Bhoomi project was judged to be so successful that other Indian states decided to replicate it. However, problems arose after implementation that had not been foreseen.

In many parts of rural India, a lower-level functionary is often the only government representative with whom citizens interact. Bhoomi eliminated these functionaries in the name of efficiency. Shifting citizens' access from lower (village) level to higher (state) level officials may provide freedom from the fraudulent practices that lower level officials have so often been discredited with, but it exposes the villager to more complex state processes at a higher level. The effect that this may have on citizens' freedom is not obvious. "Most of the Indian development planners are still fixated on increase in national incomes and they now also want the state to play a minimal role, in line with what some influential international donor agencies would recommend, and this gets reflected in their technology design choices. This is particularly true of the Bhoomi project of Karnataka, where ICT use is directed for improving government service delivery by reducing the role of lower-level government functionaries [...]. The desire for technical solutions to development problems should not take on a life of its own where we forget that development is about people and what they think and how they feel matters." ([94], page 276).

12.5.6   Future developments

Researchers and seasoned practitioners alike agree on the way forward: bottom-up, incremental implementation, with enough latitude to make adjustments to original plans and fix new, unforeseen problems, may be key to the success of large government informatization programmes such as DURP and Bhoomi ([20, 108, 94, 116]). These programmes never start from scratch. They wrestle with the inertia of the "installed base"—old processes, old tasks, old organized practices—and inherit the strengths and limitations of that base. Bottom-up strategies need to link the old and the new in interoperable ways. The "old" significantly influences how the "new" can be designed and how new practices can be scaled up. A "cultivation" approach—emphasizing practices already in place—may be wiser than a "construction" approach that focuses on centralized, top-down planning. Implementation of large informatization programmes in the North and the South is not a well-defined process with pre-configured beginning and end states, but an ongoing process of ecological change, characterized by "unanticipated effects" and "drift", reflecting our all-too-human inability to fully anticipate future events.

12.6 Participatory use of GIS

12.6.1 Introduction to participatory GIS

When discussing route planning in Section 12.1, we saw many different considerations to be taken into account when deciding upon how to get from point A to point B. One road may lead through an urban development with traffic, factories and landfills, while another may pass through a forest and meadows. Which route would a user prefer? Would she or he prefer the forest road even if it is 20% longer? Or maybe even if it is 50% longer? Is it a daily commute to work, or is it a day's outing? All these factors matter.

Many more considerations come into play when a user aims not just to choose the best route, but when that user actually has to build a new road, where hundreds of jobs and millions of euros are at stake, and with an impact that will last for centuries to come. Similar concerns arise for any major project that can affect people, habitat or ecosystems. In many countries, such projects require an Environmental Impact Assessment (EIA)—a legally binding procedure of assessment of all possible environmental, social and economic outcomes of the proposed activity. The purpose is to make sure that no important consequence of a project is omitted. The importance of EIAs has been recognized for several decades. The International Association for Impact Assessment (IAIA: `http://www.iaia.org/`) is a global network of over 1600 members working to promote best practices in the use of impact assessment for informed decision-making about policies, programmes, plans and projects. IAIA defines EIA as the "process of identifying, predicting, evaluating and mitigating the biophysical, social, and other relevant effects of development proposals prior to major decisions being taken and commitments made".

12.6.2 Requirements for participatory GIS

The Netherlands Commission for Environmental Assessment (NCEA: `http://www.eia.nl/`) emphasizes that public participation is key to the EIA process. This process should record the impact, alternatives and comments from the public in a report, which should be binding when making the final decision, and the public should be informed about that decision. In European Directive 2001/42/EC (also known as the Strategic Environmental Assessment (SEA) Directive) "on the assessment of the effects of certain policies, plans and programs on the environment" requires a formal environmental assessment of all activities that are likely to have significant effects on the environment. Authorities that develop and/or adopt such an activity must prepare a report on the likely significant environmental effects of that activity, consult with environmental authorities and the public, and take the report and the results of the consultation into account during the preparation process and final decision-making for the plan or programme.

The Aarhus Convention (`http://www.unece.org/env/pp/welcome.html`) takes the process one step further, linking environmental rights and human rights. The convention acknowledges that we have an obligation to future generations. It stipulates stakeholder involvement as a prerequisite of sustainable development, links government accountability and environmental protection, and focuses on interactions between the public and public authorities in a democratic context. It gives the public open access to information about development projects and relevant data.

Since EIA/SEA is a stakeholder-driven process, it sets additional requirements on how research is conducted and how results are delivered. The assessment will be successful and actually used only if it is open and transparent. As we may be dealing with mixed interests and priorities, which can easily result in disputes and even court hearings, it

is important that data, methods, analytical tools and results are open to scrutiny, well documented, and reviewed. This immediately requires that the study be scientifically rigourous and defendable. Ideally, the analysis should be peer reviewed and adopted by the community. In reality this may be hard to achieve, because of time constraints and the increasing variety and complexity of projects. To a certain extent this can be compensated for by making the study adaptive, iterative and interactive. By establishing and maintaining the assessment as an on-going open process that can be reinitiated when new data, specifications, concerns or priorities arrive, we can compensate for the inevitable uncertainty and imperfection of existing analyses. Obviously, at any stage of the project it is always important that all the deliverables are easy to interpret, understand and visualize.

12.6.3   Participatory GIS in EIA: The case of the Via Baltica highway

So how can EIA be performed and how can spatial information help? Suppose we are considering building a new highway, as was the case in Poland where the Via Baltica highway was proposed as a major improvement for accessibility between the EU's Central European nations (Figure 12.26), only to be suspended in 2007 owing to fears of irreversible ecological damage to important nature sites protected under EU law. Apparently, although several economically and environmentally more-sound alternative routes existed, they had never been considered as acceptable alternatives by the decision-makers. A systemic, spatial method for generating efficient transportation alternatives should take into account environmental regulations and concerns. At the same time it should integrate equally important considerations such as transport system efficiency, safety, socio-economic demands, and technical and financial viability as well as supporting stakeholder involvement throughout the whole planning process. To address this need, a participatory GIS interface was developed using spatial multi-criteria evaluation and network analysis to enable an objective comparison of various route alternatives [57].

Assessment criteria reflect stakeholder concerns and a wide variety of impacts arising from infrastructural development. For the Via Baltica project, representatives of environmental NGOs, Polish government bodies, independent research institutes and universities, as well as private consultants, were approached to provide a list of criteria relevant to the project. A wide range of criteria were considered, such as current traffic densities, national and landscape parks (and reserves) that should be protected, proximity to wetlands and peat bogs, proximity to urban areas, proximity to hazardous areas, risk of accidents in urban areas, location of highly fertile agricultural soils, intersections with water bodies, intersections with secondary roads, and proximity to economic zones. These criteria were grouped according to overall sustainable development objectives into four main themes: (1) transport efficiency, (2) ecology, (3) social impact and safety, and (4) economic costs and benefits.

Clearly, for different stakeholders the importance of each of these criteria was different. Each group of stakeholders was asked to give their preferences and as a result four different policy visions were defined by assigning different weights to each of the criteria themes (Table 12.2). One vision reflects equal importance of the criteria while the other visions reflect stakeholder opinions.

With these policy visions in mind, a series of suitability maps was generated. Spatial Multi-Criteria Analysis (SMCA) is a process in which the geospatial data sets representing the different criteria and weights described above are combined to prepare a routing suitability map for each of the four policy visions (Figure 12.27). Such a suitability map provides a continuous geographic surface, with each pixel value of this surface indicating the overall suitability value for routing the highway through that
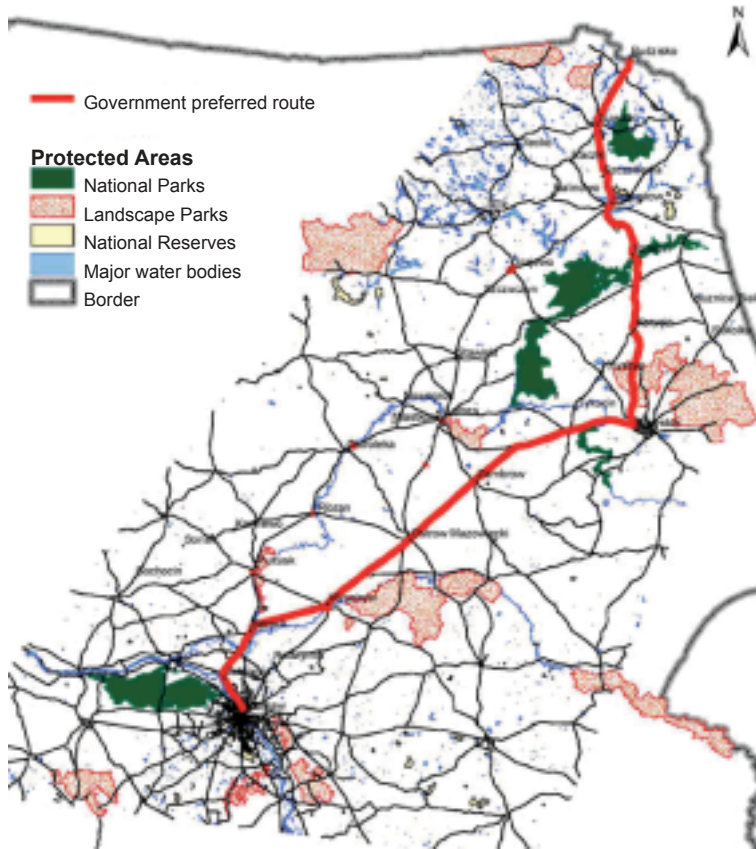
Figure 12.26
Part of the Polish road
network with indicated in red
the alternative preferred by
the Polish government.

Table 12.2
Different visions, themes and
their weights as used in the
Via Baltica corridor study.

Themes	Policy visions			
	Equal	Social	Ecology	Economy
Transport efficiency	0.25	0.27	0.27	0.27
Ecology	0.25	0.06	0.52	0.06
Social impact & safety	0.25	0.52	0.15	0.15
Economic costs & benefits	0.25	0.15	0.06	0.52

pixel.

Next, based on the suitability maps of the four visions a network of existing roads was brought into a GIS. In the Via Baltica study, four optimal routes were generated, one for each of the policy visions. It was shown that all four optimal routings had less impedance and were also shorter than the Polish government's preferred route (see Figure 12.28). In addition, each of these alternative routings would also satisfy the EU's environmental laws and enjoy a high degree of stakeholder satisfaction.

The developed methodology provides decision-makers with a tool that enables them to make more rational and transparent decisions. However, it is not always clear how to make such decisions and how to bring together the conflicting values and priorities of different stakeholders. Environmental NGOs will be unhappy with any alternative other than one based on ecological priorities. Local governments will be promoting so-
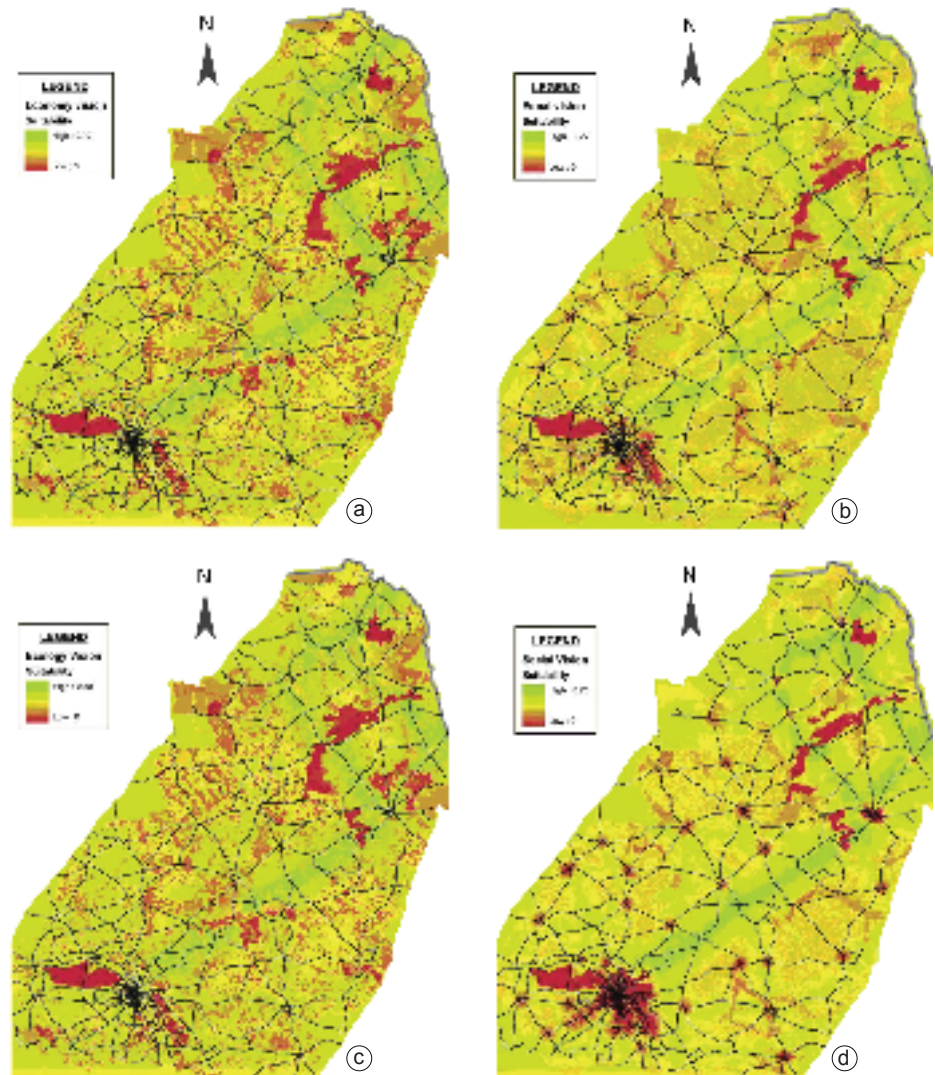
Figure 12.27
Suitability maps for the Via
Baltica route according the
four policy visions.

cietal values, while a national government will most likely be concerned with overall
economic efficiency. How can these different, possibly conflicting, views and opinions
be reconciled?
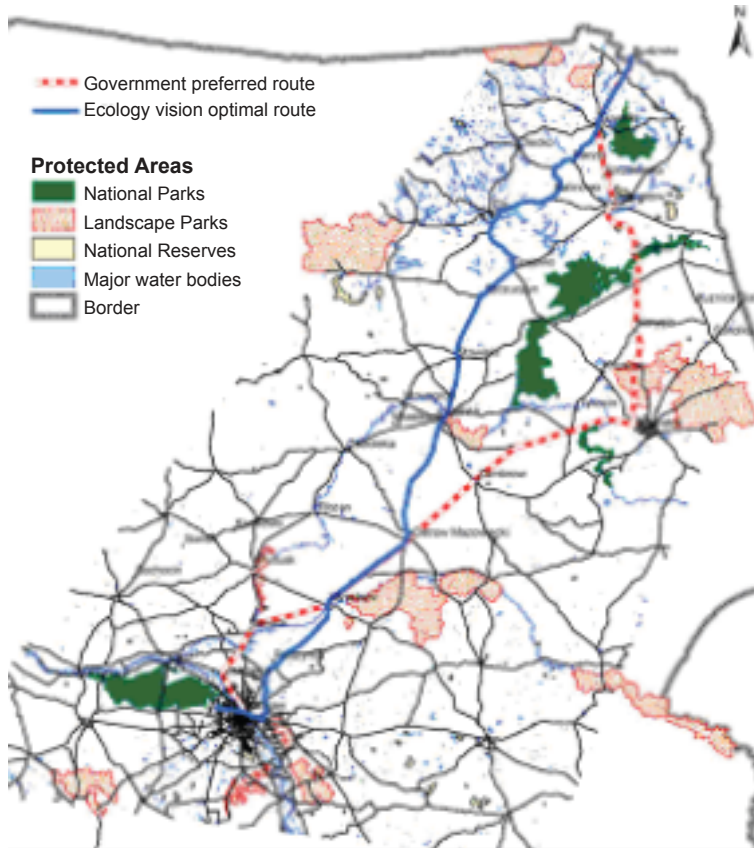
Figure 12.28
Route preferred by the Polish government (red) versus the ecology vision route (blue).

From EU Birdlife International [9]:

Recently the Polish government decided to route the highway via Lomza, which is not only the environmental sound option but also valid on economic, traffic and societal grounds. Konstantin Kreiser, EU policy manager at the BirdLife European Division commented: "This case shows once more that conflicts between nature conservation and infrastructure development can be solved through proper planning and political will—unfortunately it took the authorities seven years to learn this lesson as far as Via Baltica is concerned".

Dr Helen Byron, a senior RSPB international site casework officer, added: "Sadly, this doesn't mean our work is over entirely—we still need to protect sites along the "old" Via Baltica route and ensure that construction on the new route goes ahead so that this isn't just a paper victory. But this is an absolutely fantastic step forward, ensuring a brighter future for the wildlife of this naturally diverse region".

An approach that can help communicate information between disparate and possibly conflicting groups of stakeholders is known as *participatory modelling*. Models of all kinds—spatial and non-spatial, dynamic and static, quantitative and qualitative—

participatory modelling

always have been used in the EIA process. In most cases, however, modelling is conducted by a group of professionals, and if stakeholders are involved they are brought in either as a source of information (opinions, data, expertise) or to peruse modelling results. A model that is mainly developed externally to the process is, therefore, often treated with distrust and suspicion. Model uncertainties and insufficiencies can be easily "blown up" and exaggerated, especially if there are controversial opinions or if "inconvenient" decisions have to be made. The recent debate on climate change exemplifies this clearly: no action was taken for a long time under the pretext that model results were "uncertain".

At the same time, as anybody ever involved in modelling would confirm, a model itself is an extremely powerful tool for learning and understanding. When a model is built, the system has to be carefully analysed, the key data need to be considered, and all the most important links and connections need to be evaluated. Participatory modelling, with its various types and clones, has emerged as a powerful tool that can (a) enhance stakeholders' knowledge and understanding of a system and its dynamics under various conditions, as in collaborative learning, and (b) identify and clarify the impacts of solutions to a given problem, usually related to supporting decision-making, policy, regulation or management.

Let us consider the example of Lake Champlain in Vermont, U.S.A., which has been receiving excess-nutrient runoff for the past 50 years owing to changes in agricultural practices and rapid development of open space for residential use (Figure 12.29). The effect of excess nutrients has been most dramatically witnessed in bays such as St. Albans Bay, which became eutrophic and turned green from algal blooms every August [46].

The watershed feeding St. Albans Bay is dominated by agriculture, but it is also affected by population growth in surrounding urban areas. In the 1980s, urban point sources of pollution were reduced by upgrading the St. Alban's sewage treatment plant and, at the same time, agricultural non-point pollution was addressed through implementation of "best management practices" (BMPs) on 60% of the farms in the watershed. The cost was US $2.2 million [113].

Despite the considerable amount of money and attention paid to phosphorus loading in St. Albans Bay, it still remains a problem to this today. The focus has remained primarily on agricultural sources in the watershed and, as a result, this has caused considerable tension between farmers, city dwellers, and landowners with lake-front property. Residents blame farmers for applying too much fertilizer and manure, and farmers blame residential dwellers for maintaining too many lavish lawns, the untreated storm water and the creation of impervious surfaces. As always, the brewing discontent is exacerbated by regulations. The Lake Champlain total maximum daily load (TMDL) of phosphorus allocated to the St. Albans Bay watershed required a 33% reduction of total phosphorus flows to the bay [84]. How should this reduction be shared between farming communities and urban residents?

This was not an easy problem to solve because it concerned non-point source pollution generated over a vast watershed that was then transported to the lake along a variety of pathways. Statistical, mass-balance and dynamic landscape simulation models were used to assess the state of the watershed, and the long-term accumulation of phosphorus in it, and to describe the distribution of the average annual phosphorus load to streams in terms of space, time and transport processes. A participatory modelling effort was conducted to apportion the total load of phosphorus among all sources, including their diffuse transport pathways, and to identify the most cost-effective interventions needed to achieve target reductions. A group of stakeholders were invited to participate in the two-year research process and were engaged in the
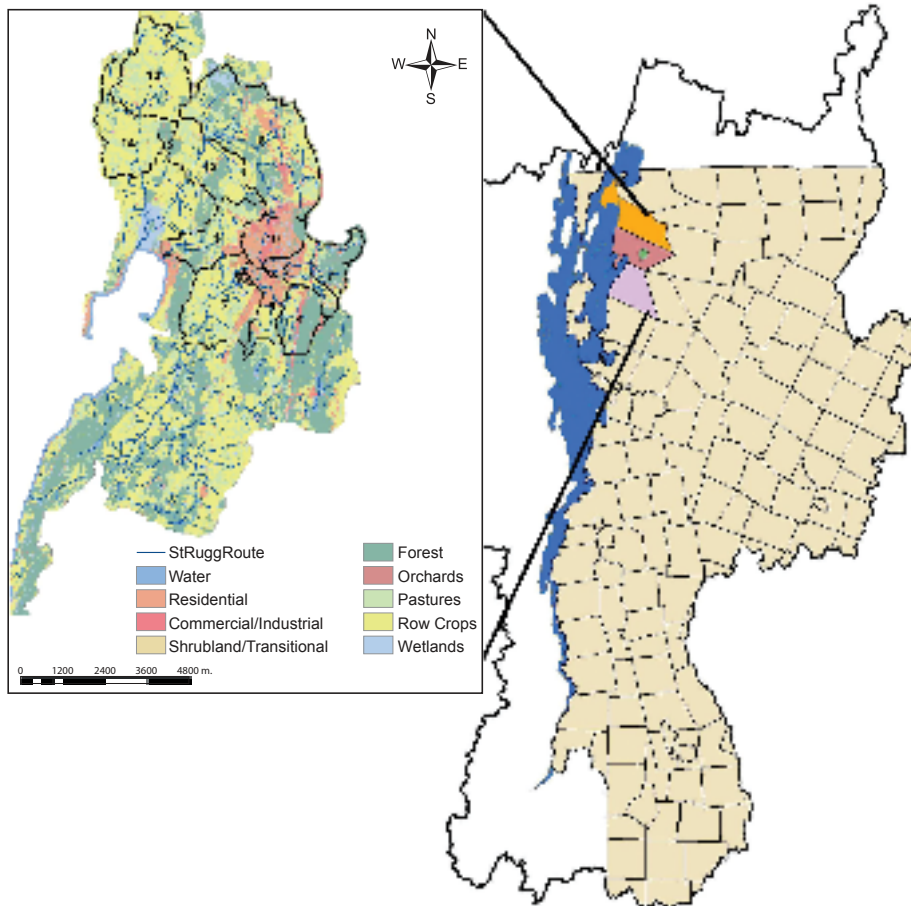
research at multiple levels, including water quality monitoring, soil phosphorus sampling, model development, scenario analysis, and future policy development [34].
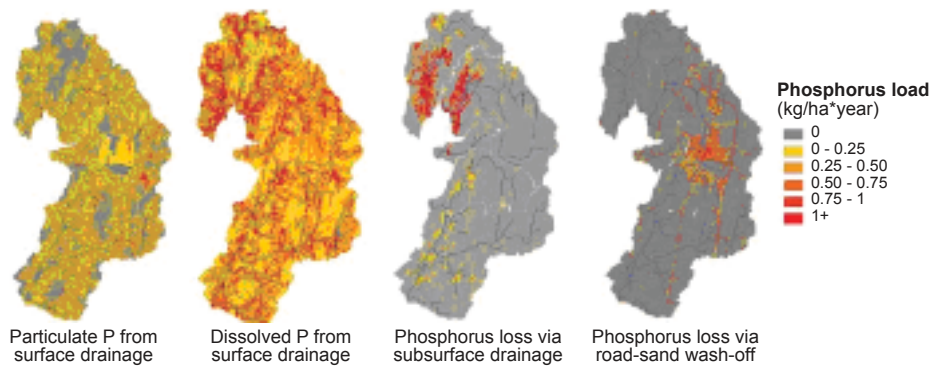
The participatory process started with a "Hydrology 101" crash course, in which stakeholders were introduced to the concepts of watershed hydrology and learned what factors can impact water quality and quantity. After that a fairly complex Landscape Modelling Framework [129] linked dynamic local models into a spatial grid of cells to describe how water and constituents are moving across space (Figure 12.30).

Watershed interventions that matched to most significant phosphorus sources and transport processes were identified with input from stakeholders and evaluated with the landscape model (Figure 12.31). Stakeholders were then invited to identify various interventions and formulate scenarios for future watershed development. Some 18 scenarios were identified, with the stakeholders having the best knowledge of what would be feasible for this particular area. The model was then used to compare the effect of various management scenarios on phosphorus loading (Figure 12.32).

Modelling results suggested that the St. Albans Bay watershed accumulates phosphorus over the long-term, primarily in agricultural soils. Dissolved phosphorus in surface runoff from the agricultural landscape, driven by high soil-phosphorus concentrations, accounts for 41% of the total load to watershed streams. Direct discharge from farmsteads and storm water drains, primarily from road-sand wash-off, were
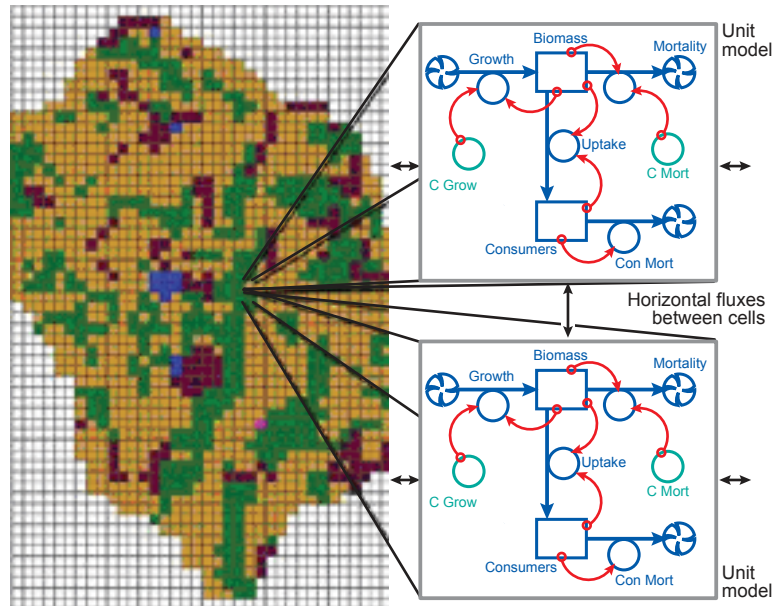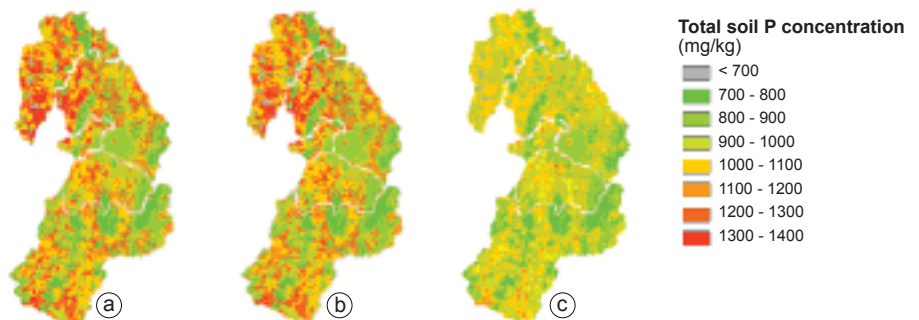
Figure 12.30
A raster map of the St. Albans Bay watershed with local processes described by process-based models.

Figure 12.31
Sources of phosphorus loading identified from the spatial landscape model.

Particulate P from surface drainage

Dissolved P from surface drainage

Phosphorus loss via subsurface drainage

Phosphorus loss via road-sand wash-off

Phosphorus load
(kg/ha*year)
- 0
- 0 - 0.25
- 0.25 - 0.50
- 0.50 - 0.75
- 0.75 - 1
- 1+

also found to be significant sources. Spatial optimization algorithms were then applied to identify the best mix of interventions at different locations to improve the overall efficiency, while minimizing total costs.

Figure 12.32
15-year scenario runs with the model. a) Base run; b) Fertilizer elimination scenario; c) Fertilizer elimination and reduced manure scenario.

Total soil P concentration
(mg/kg)
- < 700
- 700 - 800
- 800 - 900
- 900 - 1000
- 1000 - 1100
- 1100 - 1200
- 1200 - 1300
- 1300 - 1400

The participatory modelling approach used in this study led to the identification of

different solutions than stakeholders had previously assumed would be required. The approach also led to greater community acceptance and use of the modelling results, as demonstrated by local decision-makers being prepared to implement some of the solutions identified to be most cost-effective.

One of the keys to the success in this project, as with any participatory approach, is that the community participating in the research was consulted from the very beginning of the project and was urged to set project goals and identify the specific issues to be studied. In the St. Albans project stakeholders were engaged in the decision-making process through knowledge provision, model selection and development, data collection and integration, scenario development, interpretation of results, and development of policy alternatives. Engaging participants in as many of these phases as possible and as early as possible—beginning with setting the goals for the project—drastically improves the value of the resulting model in terms of its usefulness to decision-makers, its educational potential for the public and its credibility within the community.

From stakeholder interviews in the St.Albans project [34]:	"It brought people together in a non-confrontational manner. That was key, I thought. It wasn't that we were coming together to discuss what the farmers were doing, or what the city folk were doing. We were coming together to just say this is what is happening without any blame being placed on anybody." "Tying pieces of information together was an important aspect of this process."

### 12.6.4   Future developments in participatory use of GIS

The drive towards participatory decision-making is primarily fuelled by the increasing realization that the more humans impact the environment and the more they attempt to manage natural resources, the more complex and less predictable the overall socio-ecological system becomes. And the harder it becomes to find the right decisions to be made and to choose best management practices. Participatory modelling helps to "level the playing field" for decision-making by providing a common pool of knowledge and data that are delivered in the process of shared learning by the stakeholders. Participatory modelling can also improve communication between formerly disconnected groups of stakeholders, providing a common language for interaction and resolving disputes, which leads to more consensus and easier decision-making, as well as better decisions.

The St. Albans example shows how EIA and SEA call for an elevated role for users in the decision-making process. In this case, by engaging users in the process rather than bringing them in only to receive the end-product, the standard flow of delivery was changed. For EIA, we need user input upfront: we need them to define the scope of the study and the main goals, issues and scenarios that need to be analysed. Moreover, by keeping users engaged in a participatory modelling effort we ensure their "buy-in" for the results of the study, we help them better understand the issues and the various links and value sets of other, possibly conflicting, parties. This can help user communities to resolve disputes and make better, mutually acceptable decisions.

12.7   Concluding remarks on users and user requirements

Whether it concerns land or sea, demand for GI Science has always been high. GI Science products and services have become easier and cheaper to access. And as their usability improves, we have seen an ever-increasing interest and rapid uptake of the technology by many users. What was once exclusively the domain of scientists and engineers is becoming accessible to non-specialists who wish to navigate easily between locations or access, analyse and use spatially and temporally referenced information. The range of products, services and data described in this chapter highlight the breadth and depth of information available and the delivery mechanisms through which those services and information can be accessed.

Technology is rapidly developing. Even a technophobe will be aware that developments in science and engineering are changing how we interact, discuss, make decisions and act. Most information used for decision-making, governance, design and debate is referenced in space and time. As such, GI Science is becoming readily accessible, and its availability is still increasing.
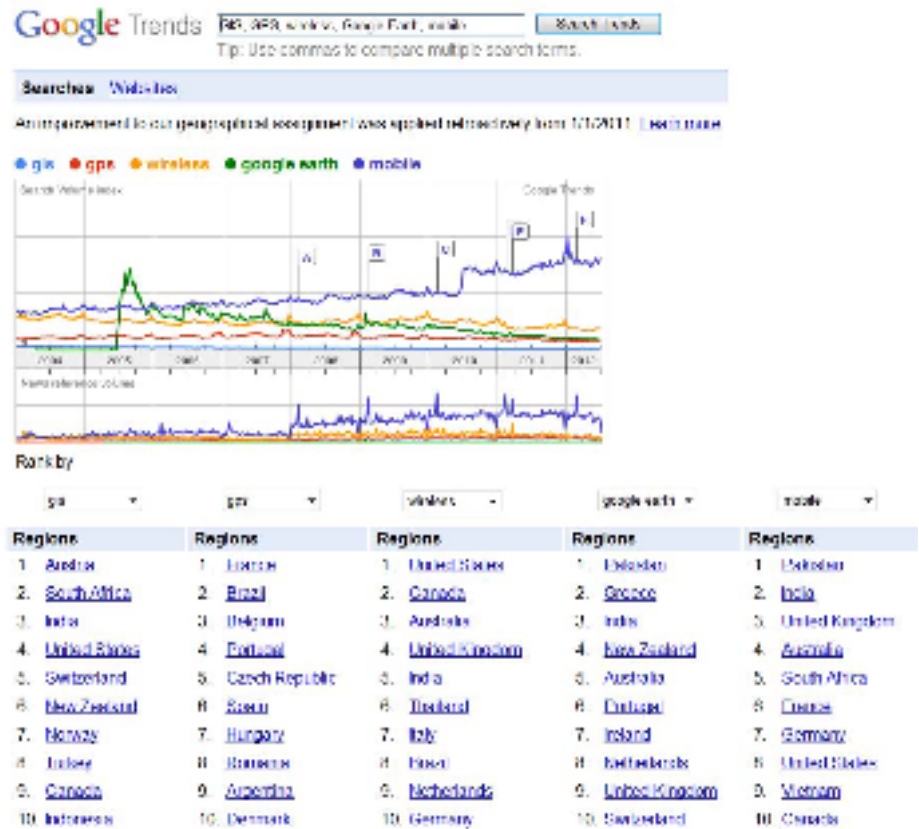
Figure 12.33
Google trends. Screenshots of Google trends combined in an analysis of keywords.

Two technologies that impact our lives, and are central to GI Science, are the Internet and wireless (mobile) communication. As a Google trends analysis indicates (Figure 12.33), while GISs are of relatively low interest to the Internet community and the news media, the number of times the term *mobile* is searched across the web is increasing. New technologies such as Google Earth tend to initially peak and then stabilize. Practical GIS tools, such as GPS, are more frequently researched by the Internet com-

munity and the news media. The combination of Internet and wireless/mobile technology enables users to easily access GI Science applications (think of Google Earth on mobile devices). Real-time updating of GI Science applications (e.g. updates on traffic flows while driving, see Section 12.1) is another example of a clever combination of wireless technology and GISs.

Information and warnings about possible or impending natural disasters such as floods may be accessed on the Internet via mobile devices (Section 12.2), while *hard* coastal defence systems may be activated (e.g. storm-surge barriers) and vegetation monitored using RS, resulting in action taken to strengthen *soft* defence systems (Section 12.3).

GI Science offers tools to assist in complex decision-making in society (such as spatial planning or environmental impact assessment for a newly proposed development as described in Sections 12.4 and 12.5). We have illustrated the importance of:

- providing information to stakeholders;

- creating a platform for interaction and discussion about possible scenarios and alternatives;

- detailing concrete outcomes resulting from proposed actions; and

- allowing users to participate in decision-making processes.

Where the future will take users of GI Science is difficult to gauge, but it will be on the one hand defined by users, by their needs and their priorities and on the other hand by technological development. Tools to observe and sense the environment will become smaller, cheaper and ubiquitous, generating volumes of online data that will dwarf the volume of data and information currently available. What is even more important is the new methods and analytical tools of the future that will allow users to "wrap" this information into useful and meaningful outcomes that will be easy to visualize, interpret and understand. Here, GI Science plays the critical role to generate information out of the data, and to generate downstream products out of primary products. Development of methods to do so is crucial to remaining "afloat" and being able to "swim" in this rising sea of information.

Wireless (mobile) technologies will continue to improve in speed and capacity. Georeferencing of personal information by place and date is generating huge databases of spatial information that are available to central governments and large IT companies such as Google and Microsoft. Almost everything can be mapped nowadays, whether it concerns medical records, travel movements, financial transactions, monitoring by surveillance cameras in public places or along roads, or simply the location of one's mobile phone. The geo-component of data, and innovative ways of analysing and visualizing these data, and then turning them into information, gives much scope for further research. A new and fascinating development may quickly become standard technology. An example is mapping the digital relationships between all of someone's friends on Facebook via the touchgraph map (see Figure 12.34). Friends who have a very close connection to you are closer to the centre of the graph than those who have fewer connections with you. And, in general, those who are more connected are relatively close together in the graph while others who are hardly or not in contact are at opposite ends of the graph. These networks will most likely further develop and become the key of new uses of GI Science.

Figure 12.34
A touchgraph facebook map.
Source: [50].

powered by
TouchGraph

Bibliography

[1] P.C. Annez and R.M. Buckley. *Urbanization and growth*, chapter Urbanization and growth: setting the context. World Bank, Washington, 2009.

[2] G. Arbia, D. Griffith, and R. Haining. Error propagation modelling in raster gis: overlay operations. *International Journal of Geographical Information Science*, 12(2):145–167, 1998.

[3] O. Arino, D. Fernandez-Prieto, and E. Volden. Healing the Earth, Earth observation supporting international environmental conventions. *ESA Bull.*, 128:53–60, 2006.

[4] S. Aronoff. *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa, Canada, 1989.

[5] J.E. Avendano Castillo. Route optimization for hazardous materials transport. MSc thesis, 2004.

[6] O. Barreteau, C. LePage, and P. Perez. Contribution of simulation and gaming to natural resource management issues: an introduction. *Simulation&Gaming*, 38:185–194, 2007.

[7] J. Bertin. *Sémiology Graphique*. Mouton, Den Haag, The Netherlands, 1967.

[8] W. Bijker. *Radar for rain forest—A monitoring system for land cover change in the Colombian Amazon*. PhD thesis, International Institute for Aerospace Survey and Earth Sciences, Enschede, The Netherlands, 1997.

[9] BirdLife International. Via baltica—another landmark victory for poland's nature. `http://www.birdlife.org/news/news/2009/10/via_baltica.html`, 2009.

[10] C. Board. Report of the working group on cartographic definitions. *Cartographic Journal*, 29(1):65–69, 1990.

[11] F. Bousquet, O. Barreteau, C. LePage, C. Mullon, and J. Weber. An environmental modelling approach. the use of multi-agents simulations. In F. Blasco and A. Weill, editors, *Advances in Environmental and Ecological Modelling*, pages pp. 113–122. Elsevier, Paris, 1999.

[12] S. Brocklesby, J.and Cummings. Foucault plays habermas: An alternative philosophical underpinning for critical systems thinking. *The Journal of the Operational Research Society*, 47:741–754, 1996.

[13] P.A. Burrough and A.U. Frank. *Geographic Objects with Indeterminate Boundaries*. GISDATA Series. Taylor & Francis, London, 1996.

[14] F. Capra. *The Tao of physics: An Exploration of the Parallels Between Modern Physics and Eastern Mysticism*. Shambhala Publications, Boston, 1975.

[15] F. Capra. *The Web of Life: A New Scientific Understanding of Living Systems*. Anchor Books, New York, 1997.

[16] A.P. Carleer and E. Wolff. VHR image region-based classification potential in the framework of the control with remote sensing of the European CAP. In *Proceedings SPIE*, volume 5976, 2005.

[17] CBS. Central bureau for statistics. `http://www.cbs.nl`, 2012.

[18] N. R. Chrisman. Errors in categorical maps: testing versus simulation. In *Proceedings AutoCarto*, pages 521–529, 1989.

[19] M. Christie. Data collection and the ozone hole: Too much of a good thing? In *Proceedings of the International Commission on History of Meteorology*, volume 1, pages 99–105, Mexico City, Mexico, 2004.

[20] C.U. Ciborra. *From control to drift: the dynamics of corporate information infastructures*. Oxford University Press, Oxford, 2000.

[21] R.N. Clark. Spectroscopy of rocks, and minerals,, and principles of spectroscopy. In A.N. Rencz, editor, *Manual of Remote Sensing: Remote Sensing for the Earth Sciences*, volume 3, pages 3–58, New York, 1999. John Wiley & Sons.

[22] D. G. Clarke and M. Clark. Lineage. In S. C. Guptill and J. L. Morrison, editors, *Elements of Spatial Data Quality*, pages 13–30. Elsevier Science, Oxford, U.K., 1995.

[23] Commissie Bodemdaling, Jaarverslag 25, 2008.

[24] CRED. EM-DAT: The OFDA/CRED International Disaster Database. `http://www.emdat.be/`, 2012.

[25] F. et al. Danielsen. The asian tsunami: A protective role for coastal vegetation. *Science*, 310:643, 2005.

[26] I. Delikostidis and C.P.J.M. van Elzakker. *In: Location based services and telecartography II : from sensor fusion to context models : 5th International Conference on Location Based Services and TeleCartography*, chapter Geo - identification and pedestrian navigation with geo - mobile applications : how do users proceed?, pages 185–206. Lecture Notes in Geoinformation and Cartography. Springer, Berlin, 2009.

[27] Deltawerken online. `http://http://www.deltawerken.com/`.

[28] R. Devillers and R. Jeeansoulin. *Data Structures and Algorithms*. ISTE Ltd, London, United Kingdom, 2006.

[29] D. DiBiase. Visualization in Earth Sciences. *Earth and Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences*, 59(2):13–18, 1990.

[30] A. Ebert, N. Kerle, and A. Stein. Urban social vulnerability assessment with physical proxies and spatial metrics derived from air- and spaceborne imagery and gis data. *Natural Hazards*, 48:275–294, 2009.

[31] J.C. Farman, B.G. Gardiner, and J.D. Shanklin. Large losses of total ozone in Antarctica reveal seasonal C10x/NOx interaction. *Nature*, 315:207–210, 1985.

[32] J. Fiebleman. Theory of integrative levels. *British Journal for the Philosophy of Science*, 5:59–66, 1954.

[33] T. Filatova, A. Voinov, and A. van der Veen. Land Market Mechanisms for Preservation of Coastal Ecosystems: an Agent-Based Analysis. *Environmental Modelling & Software*, 26:179–190, 2011.

[34] E. Gaddis, H.H. Falk, C. Ginger, and A. Voinov. Effectiveness of a participatory modeling effort to identify and advance community water resource goals in St. Albans, Vermont. *Environmental Modelling & Software*, 25:1428–1438, 2010.

[35] Geo-standaarden (geo-standards), in dutch, 2009. `http://www.geonovum.nl/geostandaarden`.

[36] P.Y. Georgiadou and J.E. Stoter. Studying the use of geo—information in government: a conceptual framework. *Computers, Environment and Urban Systems*, page 9, 2009. IN PRESS.

[37] Geoss 10 year implementation plan. `http://www.earthobservations.org/geoss_imp.shtml`.

[38] Gisdesk radboud university nijmegen. `http://www.ru.nl/gisdesk/geo-data/algemeen_beschikbaar/`.

[39] M. Goodchild. Geographical information science. *Int. J. Geographical Information Systems*, 6(1):31–45, 1992.

[40] Google Maps. `http://maps.google.com/`.

[41] S.M.E. Groten. Land ecology—and land use survey. ITC Lecture Notes RUS10, 1994.

[42] V. Gupta. Remote sensing and photogrammetry in treaty verification: present challenges and prospects for the future. *The Photogrammetric Record*, 14:729–745, 1994.

[43] C. Hall and J. Day. *Ecosystem Modeling in Theory and Practice. An Introduction with Case Histories*. John Wiley & Sons, New York, NY, 1977.

[44] G.B.M. Heuvelink. *Error propagation in quantitative spatial modelling—Applications in Geographical Information Systems*. Nederlandse Geografische Studies. Koninklijk Aardrijkskundig Genootschap, Utrecht, 1993.

[45] M. Hootsmans and H. Kampf. Ecological networks: Experiences in the netherlands. Technical report, Ministry of Agriculture, Nature and Food Quality Reference Centre (Expertisecentrum-LNV), 2004. `http://www.minlnv.nl`.

[46] K. Hyde, N. Kamman, and E. Smeltzer. History of phosphorus loading to st. albans bay, 1850–1990. Technical Report Technical Report No. 7B, Lake Champlain Basin Program, 1994.

[47] J. Iliffe. *Datums and Map Projections for Remote Sensing, GIS and Surveying*. Whittles Publishing, CRC Press, 2000.

[48] The (iso/tc 211) technical committee on geographic information/geomatics. `http://www.isotc211.org/,` accessed on October 2009.

[49] V.G. Jetten, A.P.J. De Roo, and D. Favis-Mortlock. Evaluation of field-scale and catchment-scale soil erosion models. *Catena*, 37:521–541, 1999.

[50] Jonyang.org, 2009. `http://www.jonyang.org/2009/03/everyone-else-and-you.html.`

[51] A.G. Journel and C.J. Huijbregts. *Mining geostatistics*. Academic Press, London, 1978.

[52] Wolfgang Kainz. Logical consistency. In S. C. Guptill and J. L. Morrison, editors, *Elements of Spatial Data Quality*, pages 109–137. Elsevier Science, Oxford, U.K., 1995.

[53] N. Kerle and D. Alkema. *Applied urban ecology : a global framework*, chapter Multiscale Flood Risk Assessment in Urban Areas - A Geoinformatics Approach, pages 93–105. Wiley-Blackwell, 2012.

[54] N. Kerle, S. Heuel, and N. Pfeifer. *Geospatial Information Technology for Emergency Response*, chapter Real-time data collection and information generation using airborne sensors, pages 43–74. Taylor & Francis, London, 2008.

[55] N. Kerle and C. Oppenheimer. Satellite remote sensing as a tool in lahar disaster management. *Disasters*, 26:140–160, 2002.

[56] A. M. Kerr, A. H. Baird, and S. J. Campbell. Comments on "Coastal mangrove forests mitigated tsunami" by K. Kathiresan and N. Rajendran [Estuarine, Coastal and Shelf Science, 65:601-606,2005]. *Estuarine, Coastal and Shelf Science*, 67:539–541, 2006.

[57] S.S. Keshkamat, J.M. Looijen, and M.H.P. Zuidgeest. The formulation and evaluation of transport route planning alternatives: a spatial decision support system for the via baltica project, poland. *Journal of Transport Geography*, 17:54–64, 2009.

[58] H. T. Kiiveri. Assessing, representing and transmitting positional accuracy in maps. *International Journal of Geographical Information Systems*, 11(1):33–52, 1997.

[59] R.A. Knippers and J. Hendrikse. Coordinaattransformaties. *Kartografisch Tijdschrift*, 3, 2000.

[60] Menno-Jan Kraak and F. J. Ormeling. *Cartography: Visualization of Spatial Data*. Pearson Education, Harlow, U.K., 2nd edition edition, 2003.

[61] K. Kraus et al. *User Manual SCOP*. University of Technology Vienna, Austria, 1998.

[62] A. Kroon. *The physical geography of Western Europe*, chapter Marine and coastal environments, page 438. Oxford University Press, 2005.

[63] I. Kuriyama. Supporting multilateral environmental agreement with satellite Earth observation. *Space Policy*, 21:151–160, 2005.

[64] Gail Langran. *Time in Geographic Information Systems*. Technical Issues in Geographic Information Systems. Taylor & Francis, London, U.K., 1992.

[65] R. Laurini and D. Thompson. *Fundamentals of Spatial Information Systems*, volume 37 of *The APIC Series*. Academic Press, London, U.K., 1992.

[66] D. B. Lee. Requiem for large-scale models. *Journal of the American Institute of Planners*, 39:163–187, 1973.

[67] T.M. Lillesand, R.W. Kiefer, and J.W. Chipman. *Remote Sensing and Image Interpretation*. John Wiley & Sons, New York, NY, fifth edition, 2004.

[68] X.H. Liu, K. Clarke, and M. Herold. Population density and image texture: A comparison study. *Photogrammetric Engineering and Remote Sensing*, 72:187–196, 2006.

[69] Paul A. Longley, Michael F. Goodchild, David M. Maguire, and David W. Rhind, editors. *Geographical Information Systems: Principles, Techniques, Management, and Applications*, volume 1. John Wiley & Sons, New York, N.Y., second edition, 1999.

[70] D. Lu, P. Mausel, E. Brondizio, and E. Moran. Change detection techniques. *International Journal of Remote Sensing*, 25:2365–2401, 2004.

[71] K.R. McCloy. *Resource Management Information Systems*. Taylor & Francis, London, U.K., 1995.

[72] Bruce McCormick, Tomas A. DeFanti, and Maxine D. Brown (eds.). Visualization in scientific computing. *ACM SIGGRAPH Computer Graphics—Special issue*, 21(6), 1987.

[73] H. Middelkoop. Uncertainty in a GIS, a test for quantifying interpretation output. *ITC Journal*, 1990(3):225–232, 1990.

[74] N. Minar, R. Burkhart, C. Langton, and M. Askenazi. The Swarm simulation system: A toolkit for building multi-agent simulations. Working Paper 96-06-042. *Santa Fe Institute*, 1996.

[75] T. Miura, H. Yoshioka, K. Fujiwara, and Yamamoto. H. Inter-comparison of ASTER and MODIS surface reflectance and vegetation index products for synergistic applications to natural resource monitoring. *Sensors*, 8:2480–2499, 2008.

[76] M.J. Molina and F.S. Rowland. Stratospheric sink for chlorofluoromethanes: Chlorine atomc-atalysed destruction of ozone. *Nature*, 249:810–812, 1974.

[77] J. L. Morrison. Topographic mapping for the twenty-first century. In D. Rhind, editor, *Framework of the World*, pages 14–27. Geoinformation International, Cambridge, U.K., 1997.

[78] B. Naimi and A. Voinov. StellaR: A software to translate Stella models into R open-source environment. *Environmental Modelling & Software*, 38:117–118, 2012.

[79] National Mapping Division, U. S. Geological Survey. Spatial data transfer standard. Technical report, U. S. Department of the Interior, 1990.

[80] D. Nebert, editor. *Developing Spatial Data Infrastructures: The SDI Cookbook*. The Global Spatial Data Infrastructure Association (GSDI), version 2.0 edition, 2004. `http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0`.

[81] *Using geoinformation technology for the establishment of a local flood early warning system*, Proceedings of the Second International Conference of Geoinformation Technology for Natural Disaster Management and Rehabilitation, Bangkok, Thailand, 2008.

[82] I. Niemeyer and S. Nussbaum. *Change detection: The potential for nuclear safeguards*, pages 335–348. Springer, Berlin, Germany, 2006.

[83] NOAA Laboratory for Satellite Altimetry. `http://ibis.grdl.noaa.gov/SAT/SeaLevelRise/slr/map_txj1j2_sst.pdf`.

[84] VTANR NYDEC. Lake Champlain Phosphorus Total Maximum Daily Load (TMDL), 2002. `http://www.anr.state.vt.us/dec/waterq/lakes/docs/lp_lctmdl-report`.

[85] The open geospatial consortium, inc. (ogc). `http://www.opengeospatial.org/`, accessed on October 2009.

[86] S. Openshaw, M. Charlton, and S. Carver. Error propagation: a Monte Carlo simulation. In I. Masser and M. Blakemore, editors, *Handling geographical information: methodology and potential applications*, pages 78–101. Longman, Harlow, U.K., 1991.

[87] OpenStreetMap. `http://www.openstreetmap.org/`.

[88] M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden, and C.E. Hanson, editors. *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom, 2007.

[89] B.F. Pedersen. The European way of controlling area based subsidies. In *Proceedings of IEEE International Geoscience and Remote Sensing Symposium IGARSS'01*, pages 1639–1641, Sydney, NSW, Australia, 2001.

[90] E. Pennisi. How will big pictures emerge from a sea of biological data? *Science*, 309:94, 2005.

[91] N. Peter. The use of remote sensing to support the application of multilateral environmental agreements. *Space Policy*, 20:189–195, 2004.

[92] C. Pohl and J.L. van Genderen. Multisensor image fusion in remote sensing : concepts, methods and applications. *International journal of remote sensing*, 19:823–854, 1998.

[93] S. Prachansri. Analysis of soil and land cover parameters for flood hazard assessment : a case study of the nam chun watershed, phetchabun, thailand. MSc thesis, 2007.

[94] A. Prakash and R. De'. Importance of development context in ict4d projects: A study of computerization of land records in india. *Information technology & People*, 20:262–281, 2007.

[95] Franco P. Preparata and Michael I. Shamos. *Computational Geometry—An Introduction*. Springer-Verlag, New York, NY, 1985.

[96] J. Randers. *Elements of the System Dynamics Method*, chapter Guidelines for Model Conceptualization. Pegasus Communications, Waltham, MA, 1980.

[97] C. Read. *Logic, deductive and inductive*. 1898.

[98] H.W.J. Rittel and M.M. Webber. Dilemmas in a general theory of planning. *Policy Sciences*, 4:155–169, 1973.

[99] A.C. Rosenqvist, M. Finlayson, J. Lowry, and D. Taylor. The potential of long-wavelength satellite-borne radar to support implementation of the ramsar wetlands convention. *Aqua. Conserv. Mar. Freshwater Ecosyst.*, 17:229–244, 2007.

[100] ruimtelijkeplannen.nl. `http://www.ruimtelijkeplannen.nl/web-roo/`.

[101] T.L. Saaty. *Decisions making for leaders*. Lifetime Learning Publications, Belmont, 1982.

[102] Hanan Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, Ma, 1990.

[103] K.S. Schmidt and A.K. Skidmore. Spectral discrimination of vegetation types in a coastal wetland. *Remote Sensing of Environment*, 85:92–108, 2003.

[104] K.S. Schmidt, A.K. Skidmore, E.H. Kloosterman, H. van Oosten, L. Kumar, and J. Janssen. Mapping coastal vegetation using an expert system and hyperspectral imagery. *Photogrammetric Engineering and Remote Sensing*, 70:703–716, 2004.

[105] T.A. SLOCUM, R.B. Mcmaster, F.C. KESSLER, and H.H. Howard. *Thematic Cartography and Geovisualization*. Pearson Education, USA, 3rd edition, 2009.

[106] J. P. Snijder. Map projections—a working manual. Professional paper 1395, 1987.

[107] H. Solomon. Gis based surface runoff modeling and analysis of contributing factors : a case study of nam chun watershed, thailand. MSc thesis, 2005.

[108] S.L. Star and K. Ruhleder. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7:111–133, 1996.

[109] R.S. Stolarski, A.J. Krueger, M.R. Schoeberl, R.D. McPeters, P.A. Newman, and J.C. Alpert. Nimbus 7 satellite measurements of the springtime antarctic ozone decrease. *Nature*, 322:808–811, 1986.

[110] J.J. Telford, J. Cosgrave, and R. Houghton. *Joint Evaluation of the international response to the Indian Ocean tsunami: Synthesis Report.* Tsunami Evaluation Coalition, London, 2006.

[111] C. Dana Tomlin. *Geographic Information Systems and Cartographic Modeling*. Prentice Hall, Englewood Cliffs, NJ, 1990.

[112] J.W. Trevett. *Imaging Radar for Resources Surveys*. Chapman and Hall Ltd., London, U.K., 1986.

[113] St. Albans bay rural clean water program, 1991.

[114] F. van der Meer and S. De Jong. *Imaging Spectrometry: Basic Principles and Prospective Applications*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 2001.

[115] D.E. Van der Vlag. *Modeling and Visualizing Dynamic Landscape Objects and Their Qualities*. PhD thesis, Wageningen University, Wageningen, The Netherlands, 2006.

[116] P. van Teeffelen and Th. Overduin. Essay: Gemeentelijk beleid in de praktijk: De route naar excellente en duurzame publieke dienstverlening, 2009. `http://www.basis-online.nl/index.cfm/1,126,473,0,html/Essay-Gemeentelijk-beleid-in-de-praktijk`.

[117] G.J.M. Velders, S.O. Andersen, J.S. Daniel, D.W. Fahey, and M. McFarland. The importance of the montreal protocol in protecting climate. *Proc. Nat. Acad. Sci. USA*, 104:4814–4819, 2007.

[118] H. Veregin. Developing and testing of an error propagation model for GIS overlay operations. *International Journal of Geographical Information Systems*, 9(6):595–619, 1995.

[119] V.I. Vernadskii. *Biosphere*. Synergetic Press, 1986.

[120] A. Voinov. *Encyclopedia of Ecology. Ecological Modeling*, chapter Parameters. Elsevier, 2008.

[121] A. Voinov. *Systems Science and Modeling for Ecological Economics*. Academic Press, 2008.

[122] A. Voinov and E.J. Gaddis. Lessons for successful participatory watershed modeling: A perspective from modeling practitioners. *Ecological Modelling*, 216:197–207, 2008.

[123] L. von Bertalanffy. *General System Theory*. New York: Braziller., 1968.

[124] VROM. The new spatial planning act gives space, November 2007.

[125] VROM. Ro-online: dé toegangspoort voor ruimtelijke plannen (in dutch), 2007.

[126] The world wide web consortium (w3c). `http://www.w3.org/`, accessed on October 2009.

[127] A. Wehr and U. Lohr. Airborne laser scanning—an introduction and overview. *ISPRS Journal of Photogrammetry & Remote Sensing*, 54:68–82, 1999.

[128] G.J. Williams. Estimating chlorophyll content in a mangrove forest using a neighbourhood based inversion approach. MSc thesis, 2012.

[129] IDEAS workgroup. Landscape modelling framework. `http://www.likbez.com/IDEAS/lmf.html`, 2012.

[130] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[131] Y. Zhang and N. Kerle. *Geospatial Information Technology for Emergency Response*, chapter Satellite remote sensing for near-real time data collection, pages 75–102. Taylor & Francis, London, 2008.

[132] I.S. Zonneveld, H.A.M.J. van Gils, and D.C.P. Thalen. Aspects of the ITC approach to vegetation survey. *Documents Phytosociologiques*, IV, 1979.

Glossary

Acronyms & Abbreviations

2D   Two-dimensional

2.5D   Two-and-a-half-dimensional. Typically applied to (aspects of) GIS applications that view their phenomena in a two-dimensional space (a plane), where coordinates are pairs (x, y), but where some coordinates are associated also with a single elevation value z. This is different from 3D GIS because with any (x, y) coordinate pair, a 2.5D system can at most associate only one elevation. A TIN structure, for instance, is a typical 2.5D structure, as it only determines single elevation values for single locations

3D   Three-dimensional

AARS   Asian Association of Remote Sensing

AATSR   Advanced along track scanning radiometer

AC   Atmospheric correction

AHN   Actueel Hoogtebestand Nederland(in Dutch)

AIS   Airborne imaging spectrometer

ALI   Advanced land imager

ALS   Airborne laser scanning

ASAR   Advanced synthetic aperture radar

ASTER   Advanced spaceborne thermal emission and reflection radiometer

AVHRR   Advanced very high resolution radiometer

AVIRIS   Airborne visible/infrared imaging spectrometer

B&W   Black-and-white (film, photographs)

BRDF   Bidirectional reflectance distribution function

CAD   Computer aided design

CCD   Charge-coupled device

CMOS   Complementary metal oxide semiconductor

DBMS   Database Management System

DEM	Digital elevation model
DINSAR	Differential InSAR
DIY	Do it yourself
DLR	German Aerospace Centre (Deutsche Luft- und Raumfahrt)
DMC	Disaster management constellation
DN	Digital number
DPI	Dots per inch
DSM	Digital surface model
DTM	Digital terrain (relief) model
EaR	Element at risk
EM	Electromagnetic
EO	Earth observation
EOS	Earth observing system
ERS-1 (and ERS-2)	European remote-sensing (satellites)
ESA	European Space Agency
FAO	Food and Agriculture Organisation
FLI	Fluorescence line imager
FOV	Field of view
GCP	Ground control point
GDA	Geospatial data acquisition
GEOSS	Global Earth Observation System of Systems
GI	Geoinformation
GIS	Geographic information system
GPS	Global positioning system
GRC	Ground resolution cell
GSD	Ground sampling distance
HALE	High altitude long endurance
HRC	High-resolution camera
HRG	(SPOT-5) high resolution geometric
HRS	(SPOT-5) high resolution stereoscopic
ICESat	Ice, cloud, and land elevation satellite

ICT	Information and communication technology
IFOV	Instantaneous field of view
IGI	Ingenieur-Gesellschaft fi£¡r Interfaces
IHS	Intensity-hue-saturation
IMU	Inertial measuring unit
INS	Inertial navigation system
InSAR	Interferometric SAR
IP	Induced polarization
IPCC	International Panel for Climatic Change
IR	Infrared
IRS	Indian remote sensing
ISO	International Organization for Standardization
ISPRS	International Society of Photogrammetry and Remote Sensing
ISRO	Indian Space Research Organization
ISS	International space station
ITC	International Training Centre
LAC	LEISA atmospheric corrector
LEISA	linear etalon imaging spectrometer array
LIDAR	Light detection and ranging
LISS	Linear imaging self-scanning sensor
LULC	Land use/land cover
LUT	Look-up table
MDM	Minimum distance to mean
MERIS	Medium resolution imaging spectrometer
MIT	Massachusetts Institute of Technology
ML	Maximum likelihood
MODIS	Moderate-resolution imaging spectroradiometer
MSG	Meteosat second generation
MSS	MultiSpectral Scanner (Landsat)
NASA	National Aeronautics and Space Administration
NDVI	Normalized difference vegetation index

NIR	Near-infrared
NOAA	National Oceanic and Atmospheric Administration
OGC	Open Geospatial Consortium
OOA	Object-oriented analysis
OSA	Optical sensor assembly
PAN	Panchromatic
PC	Personal computer
PCA	Principal component analysis
PCC	Proportion correctly classified
PMI	Programmable multispectral imager
POS	Positioning and orientation system
ppm	Part per million
PRI	Primary data (used in "pri format" of image files)
Proba	Project for on-board autonomy
RADAR	Radio detection and ranging
RAR	Real aperture radar
RGB	Red-green-blue
RMSE	Root mean square error
RPC	Rational polynomial coefficients
RS	Remote sensing
RTM	Radiative transfer model
SAM	Spectral angle mapper
SAR	Synthetic aperture radar
SDI	Spatial Data Infrastructure
SEVIRI	Spinning enhanced visible and IR imager
SI	International system of units (Le Systï£¡me International d'Unitï£¡s)
SIS	Scanning imaging spectroradiometer
SMIRR	Shuttle multispectral infrared radiometer
SONAR	Sound navigation and ranging
SPOT	Satellite Probatoire pour l'Observation de la Terre
SQL	Structured Query Language; the query language implemented in all relational database management systems

SRTM	Shuttle radar topography mission
SWIR	Short-wave infrared
TES	Technology experiment satellite
TIN	Triangulated Irregular Network
TIR	Thermal infrared
TLS	Terrestrial laser scanning
TM/ETM	Thematic mapper/enhanced thematic mapper
TOA	Top of the atmosphere
UAV	Unmanned aerial vehicle
USGS	USA Geological Survey
UTM	Universal Transverse Mercator (map projection)
UV	Ultraviolet
VIS	Visible (bands)
VNIR	Visible and near-infrared (bands)
WiFS	Wide field sensor
WMO	World Meteorological Organization
YMC	Yellow, magenta, cyan

Terms

A/D conversion   Acronym for analogue to digital conversion; the process of sampling an analogue signal and quantifying signal strength by a number (stored as binary code).

Absorption   The process in which electromagnetic radiation is converted in an object or medium into other forms of radiation (e.g. heat, or fluorescence) or causing chemical reactions (e.g. photosynthesis).

Additive colours   The additive principle of colours is based on the three primary colours of light: red, green, blue. All three primary colours together produce white. Additive colour mixing is used, e.g. on computer screens and television sets.

Aerial camera   A camera specially designed for use in an aircraft (opposed to a terrestrial camera or spaceborne camera). In the context of this book mostly used to denote an aerial survey camera. A **survey camera** is specially designed to produce images for the purpose of surveying and mapping (also referred to as metric camera; can be aerial or terrestrial). An aerial survey camera is typically designed for being mounted in an airplane to take vertical photos/images of large format; it can be a film camera or a digital frame camera or line camera. A **frame camera** is a camera in which an entire frame or format is exposed simultaneously through a lens (with a fixed focal length).

Aerial triangulation   The process of finding the exterior orientation parameters of a block of aerial photos/images using a (very) limited number of ground control points.

Aerosol   A suspension of fine solid (dust) or liquid particles (water) in a gas (air).

Affine transformation   A 2D geometric transformation (plane-to-plane transformation) that uses six parameters to account for rotation, translation, scale change, and shearing. It defines the (linear) relationship between two coordinate systems such as an image coordinate system and a map coordinate system; see also **conformal transformation**.

Agent-Based Model   (ABM) These attempt to model processes in the form of multiple (possibly interacting) agents (which might represent individuals) using sets of decision-rules about what the agent can and cannot do. As such, a key notion is that simple behavioural rules for individual agents generate complex behaviour for the entire 'system'. Agent-based models have been developed to understand aspects of complex systems, for example by incorporating *stochastic* and or *deterministic* components.

Algorithm   A procedure for solving a mathematical problem (as of finding the greatest common divisor) in a finite number of steps.

Altitude   The elevation of a (moving) object above a reference surface, usually mean sea level.

Amplitude   In the context of this book, the maximum departure of a wave from its average value; it is a measure for the strength of an oscillatory movement. For instance, the louder a sound, the larger the amplitude of the sound wave.

Aperture   In optics, an aperture is an opening through which light is admitted; it determines how collimated the admitted rays are. For a conventional camera, the aperture is materialized by the diaphragm as part of the lens. In radar, the aperture is related to the antenna length.

Aspect   The geographical direction toward which a slope faces, measured in degrees from north, in a clockwise direction.

Atmosphere   The Earth's atmosphere is the gaseous envelope surrounding the Earth. The atmosphere is classified into several layers, but there are no discernible boundaries; it gradually becomes thinner. Three quarters of the atmosphere's mass is within 11 km of the Earth's surface. An altitude of 120 km marks the boundary where atmospheric effects become noticeable during re-entry from space.

Atmospheric scattering   The process of particles or gaseous molecules present in the atmosphere redirecting EM radiation from its original path. A well known example is Rayleigh scattering.

Atmospheric window   A spectrum portion outside the main absorption bands of the atmospheric gases that can be used for remote sensing.

Attribute   Data associated with a spatial feature or sample location, stored as a column in a database table. The name of the column should suggest what the values in that column stand for. These values are known as *attribute values*.

Attribute domain   An attribute's domain is a (possibly infinite) set of atomic values. Examples are real number values, string and date.

Attribute join   The join operator takes two input relations and produces one output relation, combining the relations based on a common attribute field.

Autocorrelation   see **spatial autocorrelation**

Azimuth   In mapping and navigation azimuth is the direction to a target with respect to north and usually expressed in degrees. In radar RS azimuth pertains to the direction of the orbit or flight path of the radar platform.

Backscatter   The microwave signal reflected by elements of an illuminated surface in the direction of the radar antenna.

Band   In the context of this book, mostly short for 'wavelength band', which stands for a limited range of the EM spectrum. A sensor is sensitive to certain 'spectral bands'; see also **spectral band**. Atmospheric absorption is characterized by 'absorption bands'. The term 'band' is also frequently used to indicate one of the digital images of a multi-band image, thus the data recorded by one of the channel of a multispectral sensor; e.g. band 3 of a Landsat image, or the green band of a SPOT image.

Base data   Spatial data prepared for different uses. Typically, large-scale topographic data at the regional or national level, as prepared by a national mapping organization. Sometimes also known as *foundation data*.

Black body   A body/object that absorbs all EM radiation that hits it.

Boolean   Boolean can refer to a Boolean data type (data type with only two values, true or false) but it can also refer to Boolean map algebra. In this case it refers to expressions that use "comparison operators" and evaluate to true or false.

Bounding box   A support function that determines the minimal rectangle that covers a line or polygon feature.

Brightness   Luminance in photometric terms: the luminous intensity of a surface in a given direction per unit of projected area (equivalent to radiance in radiometry, but expressed in candela per square metre). In the book loosely used for the degree of illumination or the state of being bright; very bright seen as white, very dark seen as black. In the section on radar, 'brightness' is used as the property of a radar image in which the observed strength of the radar reflectivity is expressed as being proportional to a digital number.

Bundle of rays   A photogrammetric term to denote the geometric construct that is obtained by connecting each point measured in an image with the perspective centre by a straight line (the imaging light ray). There is one bundle of rays for each photograph and it is used for exterior orientation or aerial triangulation.

Calibration report   In aerial photography, the manufacturer of the camera specifies the interior orientation in the form of a certificate or report. Information includes the principal distance, principal point, radial lens distortion data, and fiducial mark coordinates in the case of a film camera.

Camera   An electro-optical remote sensor without mechanical components. In its simplest form, it consists of the camera body, a lens, a focal plane array of CCDs, and a storage device. We distinguish line cameras and frame cameras, depending on whether they have (a few) single line(s) of photosensitive cells or a matrix array of cells (CCDs or CMOSs); see also **aerial camera**.

Cartography   The whole of scientific, technological and artistic activities directed to the conception, production, dissemination and use of map displays

Categorical data   See **Nominal data**

CCD   (Charge-Coupled Device) A solid state detector using silicon as semi-conductor material. Many are assembled on one chip and equipped with a shift register to transport charge and sample it at the end of a CCD line after amplification. One CCD is good for one pixel in one spectral band. A camera with three channels needs three CCDs for one ground resolution cell. CCDs are used in cameras for sensing in spectral bands with wavelengths smaller than about 1 μm. Solid state detectors that are sensitive to SWIR and even TIR radiation use other semi-conductor material and they need IR blockers to reduce noise and thus increase the dynamic range.

Centroid   Informally, a geometric object's midpoint; more formally, can be defined as the centre of the object's mass, i.e. that point at which it would balance under a homogeneously applied force like gravity.

Change detection   Comparison of spatial information recorded over the same area at different times to detect changes in the features of interest.

Change matrix   Table with "from" and "to" classes for rows and columns and frequency or probability of each "from-to" combination as element.

Channel   In the context of this book, the sensor component for a particular spectral band (detector, electrical circuit, storage). Many remote sensors also have one or more communication channels.

Check point   An additional ground point used to independently verify the degree of accuracy of a geometric transformation (e.g. georeferencing, exterior orientation, aerial triangulation, orthophoto).

Class   A group, set, or kind sharing common attributes. In image interpretation and classification, we define thematic classes of terrain features; they are variables of the nominal type. For instance, typical classes of land cover are grass, wheat, trees, bare soil, buildings, etc.

Cluster   Used in the context of image classification to indicate a concentration of observations (points in the feature space) related to a training class.

CMOS   (Complementary Metal Oxide Semiconductor) A device similar to a CCD. It also has a silicon photodiode, however, a CMOS array does not have a shift register but each cell has its own MOS switch for readout.

Collinearity   A (nonlinear) mathematical model that photogrammetric restitution is based upon. **Collinearity equations** describe the relationship among image/camera coordinates, object space coordinates, and orientation

parameters. The assumption is that the exposure station of a photograph, a ground point, and its corresponding image point location all lie along a straight line.

Colour A sensation of our visual system caused by EM radiation; a visual perception property; a phenomenon of light. Different colours are caused by different wavelengths of light. We may quantify 'colour' by hue (in the IHS colour space).

Colour film Also known as *true colour film* used in (aerial) photography. The principle of colour film is to add sensitized dyes to the silver halide. Magenta, yellow and cyan dyes are sensitive to red, green and blue light respectively.

Colour infrared film Photographic film with specific sensitivity for near-infrared wavelengths. Typically used in surveys of vegetation.

Comparison operators Operators that are used in map algebra and compare the value of the input raster cell to a specified value (e.g. $<$, $>$, $\leq$).

Concave A 2D polygon or 3D solid is said to be concave if there exists a straight line segment having its two end points in the object that does not lie entirely within the object. A terrain slope is concave, analogously, is concave if it (locally) has the shape of a concave solid. See also **convex**.

Conditional expressions Map algebra expression that will check if a condition is true or false but allows to user to specify the value that should be returned in case the condition holds, and the else in case the condition does not hold.

Cones Photoreceptor cells in the retina of the eye for sensing colour; see also **rods**.

Conformal transformation A 2D geometric transformation (plane-to-plane transformation) that uses four parameters to account for rotation, translation, and scale change. It defines the (linear) relationship between two coordinate systems such as an image coordinate system and a map coordinate system; see also **affine transformation**.

Contour line An elevation isoline. Valuable especially in map production.

Convex A 2D polygon or 3D solid is said to be convex if every straight line segment having its two end points in the object lies entirely within the object. A terrain slope is convex, analogously, is convex if it (locally) has the shape of a convex solid. See also **concave**.

Coordinates Linear or angular quantities, which designate the position that a point occupies in a given reference frame or system. The **coordinate system** is based on mathematical rules and used to measure distances and/or angles in order to identify the location of points by means of unique sets of numerical values. A coordinate system requires the definition of its origin, the orientation of the axes, and units of measurement. Plane rectangular coordinates describe the position of points with respect to a defined origin by means of two distances perpendicular to each other; the two reference lines at right angles to each other and passing trough the origin are called the **coordinate axes**.

Corner reflection In radar RS: a high backscatter as typically caused by two flat surfaces intersecting at 90 degrees and situated orthogonally to the radar incident beam.

Corner reflector The combination of two or more intersecting specular surfaces that combine to enhance the signal reflected back in the direction of the radar, e.g. houses in urban areas.

Crisp boundary A crisp boundary is a boundary that can be determined with a high level of precision.

D/A conversion Acronym for digital to analogue conversion; the process of converting a binary code to a charge (an electrical signal).

Dangle Line segment (arc), that is connected to another line at one end, but not at the other end.

Data Factual information in digital form, used as basis for calculation, reasoning by a computer, etc.

Data accuracy There are three types of accuracy, positional accuracy, attribute accuracy and temporal accuracy. Positional accuracy expressed how close the digital feature is to the actual position of that feature in reality.

Data assimilation Adjusting the variables of a process model. Involves comparison of the model predictions for certain observations of a variable with actual observations of the same variable, and finding an acceptable compromise between model predictions and observations.

Data formats Can refer tot the data type or to the file format (e.g. Shapefile).

Data model Structured design of a database, or a system for storing data.

Data standards Standards for interoperability of georesources (ISO, OGC).

Database An integrated, usually large, collection of data stored with the help of a DBMS.

Database Management System A software package that allows its users to define and use databases. Commonly abbreviated to DBMS. A generic tool, applicable to many different databases.

Database schema The design of a database laid down in definitions of the database's structure, integrity rules and operations. Stored also with the help of a DBMS.

Decision table A table guiding the raster overlay process.

Delaunay triangulation A partitioning of the plane using a given set of points as the triangles' corners that is in a sense optimal. The optimality characteristic makes the resulting triangles come out as equilateral as possible. The circle going through the three corner points of any triangle will not contain other points of the input set.

Dependence The set of relations linking the values of a regionalized variable at different locations in space. Spatial dependence usually arises from a common genesis and is used with advantage in predicting an unsampled value from neighbouring observations.

Detector   A device or medium for detecting the presence of (electromagnetic) radiation.

Deterministic   (In the context of an *application model*), a procedure or function that generates an outcome with no allowance or consideration for variation. Deterministic models are good for predicting results when the input is predictable, and the exact functioning of the 'process' is known. The opposite of *stochastic*.

Dielectric constant   Parameter that describes the electrical properties of a medium. Reflectivity of a surface and penetration of microwaves into the material are determined by this parameter.

Diffuse reflection   Reflection distributed over a wide range of directions.

Digital elevation model (DEM)   A representation of a surface in terms of elevation values that change with position. Elevation can refer to the ground surface, a soil layer, etc. According to the original definition data should be in a raster format.

Digital number (DN)   The recorded digital read-out of an electronic detector. It is the quantized sampled value of the electrical signal which is generated by the detector. The DNs correspond to photon energy incident upon the detector and radiances at the detector, but have not a meaningful physical unit. In 8 bits recording, the DNs are in the range [0, 255].

Digital surface model (DSM)   In Earth observation from air or space, often synonym for DEM, elevation standing for terrain elevation.

Digital terrain model (DTM)   A digital representation of terrain relief in terms of (X,Y,Z) coordinates and possibly additional information (on breaklines and salient points). Z usually stands for elevation, and (X,Y) for the horizontal position of a point. To the concept of a DTM it does not matter whether Z is orthometric or ellipsoidal elevation. Horizontal position can be defined by geographic coordinates or by grid coordinates in a map projection. DTM data can be given in different forms (contour lines, raster, TIN, profiles, etc).

Digitizing   Generally, any process that converts an analogue representation of a physical quantity to a digital representation, i.e., numeric values. Specifically, in the context of this book, the process of manually tracing lines or other image features to obtain coordinates of these features (a 'vector representation'). We can either digitize features (a) on hardcopy images using a digitizing tablet (2D) or analogue/analytical photogrammetric plotter (3D), or (b) on softcopy images using a standard computer screen and mouse (2D **on-screen digitizing**) or a digital photogrammetric workstation (3D). The process of converting an entire hardcopy image to a digital image (i.e., a raster representation) is referred to as scanning rather than digitizing.

Dynamic range   The ratio between the maximum level and the minimum level of intensity that can be measured, also called the signal to noise ratio of the detector; it is commonly expressed in decibels (db).

Earth observation   The process of gathering information about physical, chemical, biological, geometrical properties of our planet.

Electromagnetic radiation   Energy flux in space. EM radiation can be modelled as an electro magnetic wave or as elementary particles (photons). The measurement of reflected and emitted electromagnetic radiation is an essential aspect in Earth observation.

Electromagnetic spectrum   The range of all wavelengths, from gamma rays (10^{-12} m) up to very long radio waves (10^8 m).

Elevation   The height of a surface (e.g. the terrain surface) above a reference surface (e.g. an ellipsoid or the geoid); see also **height** and **altitude**.

Emission   The act or instance of releasing (specifically EM) radiation from a body.

Emissivity   of a material is the ratio of radiation emitted by a particular material to radiation emitted by a black-body at the same temperature. It is a measure of a material's ability to absorb and re-emit radiation. Emissivity is a dimensionless quantity.

Emittance   of a body/object quantifies how much radiation is emitted by it. Instead of 'emittance' sometimes the term **exitance** is used. **Spectral (radiant) emittance** is emittance per wavelength (quantifying the spectral distribution of emitted radiation); it is measured in $W\,m^{-2}\,\mu m^{-1}$.

Energy   A fundamental entity of nature, which exists in many forms including kinetic, potential, chemical, acoustic and electromagnetic radiation. EM radiation is flux of the energy of electromagnetic waves or photons.

Error matrix   Matrix that compares samples taken from the source to be evaluated with observations that are considered as correct (reference). The error matrix allows calculation of quality parameters such as overall accuracy, error of omission, and error of commission.

Euclidean distance   Measurement in Euclidean space, straight line distance.

Euclidean space   A space in which locations are identified by coordinates, and with which usually the standard, Pythagorean *distance* function between locations is associated. Other functions, such as *direction* and *angle*, can also be present. Euclidean space is n-dimensional, and we must make a choice of n, being 1, 2, 3 or more. The case $n = 2$ gives us the *Euclidean plane*, which is the most common Euclidean space in GIS use.

Exposure station   A term of aerial photography: the point in the flight path at which the camera exposes the film or opens the shutter for exposing a CCD frame. The exposure station has elements that define the position of the projection centre and the camera attitude.

Exterior orientation   See **orientation**

False colour infrared film   See **colour infrared film**

Feature space   The mathematical space describing the combinations of observations (DN-values in the different bands) of a multispectral or multi-band image. A single observation is defined by a feature vector.

Feature space plot   A two- or three-dimensional graph in which the observations made in different bands are plotted against each other.

Features   A set of measurable properties, e.g. of the Earth's surface. e.g. we may call the set of objects resulting from human construction work - the houses, roads, irrigation channels - cultural terrain features.

Fiducial marks   Four or eight reference markers fixed on the frame of a film survey camera and visible in each photograph. Fiducials are used to compute the transformation from pixel coordinates to image/camera coordinates.

Field of view (FOV)   The viewing range of a sensor expressed as angle, usually in degrees; sometimes referred to as swath angle, or in ALS as scan angle. The FOV is one of the factors which determine the swath width of a sensor-platform system; see also **instantaneous field of view**.

Filter   (a) Physical device to suppress an input component; e.g. a yellow filter in front of a lens of a camera (it absorbs blue light). (b) Algorithm in signal/image processing for eliminating or at least reducing an unwanted component of given data, e.g. a noise filter; see also **filtering**.

Filtering   Computational process of changing given values such that a contained component is either attenuated, amplified, or extracted (e.g. smoothing an image or sharpening an image, extracting brightness edges in an image, removing off-ground points from a DSM to obtain a DTM, etc).

Flow accumulation   Raster representation showing for each cell the number of cells that have their water flow into the cell.

Flow direction   Raster representation showing for each cell the direction in which this cell will drain.

Flying height   The vertical distance between the camera/sensor position at the time of exposure and the terrain at average elevation within the ground coverage of the photo/image taken.

Focal length   The distance between the optical centre of the lens and where the optical axis intersects the plane of critical focus of a very distant object. The focal length of a survey camera is determined in a laboratory; see also **principal distance**.

Focal plane   The plane (perpendicular to the axis of the lens) in which images of points in the object space of the camera are focused.

Foreshortening   Spatial distortion whereby terrain slopes facing the side-looking radar are mapped as having a compressed range scale relative to its appearance if the same terrain was flat.

Frequency   The reciprocal of the wave period.

Gain   of a linear filter or kernel, see **kernel**.

Geo-webservice   Mechanism designed to support interoperable machine-to-machine interaction over the web.

Geocoding   The process of transforming and resampling image data in such way that these can be used simultaneously with data that are in a specific map projection. Input for a geocoding process are image data and control points, output is a geocoded image. A specific category of geocoded images are orthophotos and orthoimages; see also **georeferencing**.

Geographic field   A geographic phenomenon that can be viewed as a—usually continuous—function in the geographic space that associates with each location a value. Continuous examples are elevation or depth, temperature, humidity, fertility, pH *et cetera*. Discrete examples are land use classifications, and soil classifications.

Geographic Information System   A software package that accommodates the entry, management, analysis and presentation of georeferenced data. It is a generic tool applicable to many different types of use (GIS applications).

Geographic object   When a geographic phenomenon is not present everywhere in the study area, and is easily distinguished (in space) and named, we regard it as an object.

Geographic phenomenon   Any man-made or natural phenomenon (that we are interested in).

Georeferencing   The process of relating an image to a specific map projection. As a result, vector data stored in this projection can for example be superimposed on the image. Input for a georeferencing process is an image and coordinates of ground control points, output are the transformation parameters (a "georeferenced image").

Geospatial data   Factual information related to location on the (surface of the) Earth.

Geovisualization   Making spatial data 'visible' by means of maps generated through interactive and dynamic software tools.

GPS   A satellite surveying method providing geodetic coordinates on the Earth surface.

Grey body   A body or material that absorbs and radiates only a certain fraction of EM radiation compared to a black body.

Grey values   Grey is a mixture of black and white. The computer converts the DNs (in the range 0 to 255) of a digital image to grey values on the monitor; 0 becomes black 255 white.

Grid   A network of regularly spaced horizontal and perpendicular lines (as for locating points on a map). We may associate (field) values with the nodes of a grid, e.g. elevation to obtain a DEM; see also **raster**.

Ground control points (GCP)   A ground point reliably identifiable in the image(s) under consideration. It has known coordinates in a map or terrain coordinate system, expressed in the units (e.g. metres, feet) of the specified coordinate system. GCPs are used for georeferencing and image orientation.

Ground range   Range between the radar antenna and an object as given by a side-looking radar image but projected onto the horizontal reference plane of the object space.

Ground resolution cell (GRC)   The area on the ground corresponding to the IFOV of a detector. The area can be elliptical or rectangular. A pixel value (DN) corresponds to the averaged radiance of the ground resolution cell. The extent of the GRC is sometimes referred to as pixel size (on the ground).

Ground sampling distance (GSD)   The distance between the centres of two adjacent resolution cells of the same sensor channel. For scanners and line cameras the GSD can be different along track and across track. The GSD equals the extent of a ground resolution cell in a well designed imaging sensor. The GSD is also referred to as pixel size (on the ground).

Ground surface   The bare Earth's surface (also referred to as "bald Earth").

Ground truth   A term that may include different types of observations and measurements performed in the field. The name is imprecise because it suggests that these are 100% accurate and reliable, whereas this may be difficult to achieve.

Heat   The quality of being hot or the condition of matter which produces the sensation warmth. It is one of the primary sensations, produced by contact with or nearness to fire or any body at a high temperature.

Height   The vertical extent of an object, the distance from the bottom to the top of something protruding above a level.

Histogram   Tabular or graphical representation showing the (absolute or relative) frequency of values of a variable. In the context of digital images it relates to the distribution of the DNs of a set of pixels.

Histogram equalization   The process used in the visualization of digital images to optimize the overall image contrast. Based on the histogram, all available grey levels or colours are distributed in such way that all occur with equal frequency in the result.

Hue   Quantification (in the IHS colour space) of what we refer to as blue, green, yellow, orange, red purple, etc.

Image   In the context of this book, the optical counterpart (pictorial representation) of an object produced by an optical device or an electronic device. An example of an image is a photograph, which is the likeliness of an object or scene recorded on photographic material. Another example is the picture produced on a computer screen or a television set. The term "remote sensing image" is frequently used to either distinguish arbitrary images on an electronic display from those originating from a sensor or to denote raw data produced by an electronic sensor, which are in fact not pictorial but arrays of digital numbers; the digital numbers are related to a property of an object or scene, such as the amount of reflected light. Similarly also the term **digital image** is commonly used for an array of digital numbers, which can readily be converted to an image on a computer screen or by a printer. It is convenient to call the result of scanning a photograph or the data produced by a digital camera 'digital images'.

Image classification   The process of assigning pixels to nominal, i.e., thematic, classes. Input is a multi-band image, output is a raster in which each cell has a (thematic) code. Image classification can be accomplished using a supervised or unsupervised approach.

Image coordinate system   A system of expressing the position of a point in an image by plane rectangular coordinates. You will find different definitions of an image coordinate system in literature. In this book, we define it as

either the row-column system of a digital image, or the x-y photo system ("frame image coordinates at the time of exposure") with the principal point as origin and the x-axis in the direction of flight or the respective fiducial mark.

Image enhancement The process of improving the visual representation of a digital image, e.g. by a histogram operation or using filters.

Image interpretation The key process in information extraction from images. The application context determines what is to be considered as information. We can use visual interpretation or computer vision techniques for recognizing features and objects of interest in an image.

Image matching The process of matching features common to two or more images (finding conjugate points or lines). Digital image matching is used for various purposes; main applications are automated DSM generation and automated image orientation.

Image processing system A computer system that is specifically designed to process digital images and to extract information from them by visualizing the data, or applying models and pattern recognition techniques.

Image sensor (or imager) A photographic camera is an imaging sensor. The term, however, is mostly used for (optical-)electronic sensors. They provide data of a scene in an image fashion in the form of a two-dimensional array of DNs for each spectral band of sensing. A single element of such a 2D array is referred to as pixel. The pixel value - the DN - is an integer number in a fixed range. The range is a power of 2, depending on how many bits are used for storing a DN. 8 bits is very common, but it can be up to 16 bits especially for thermal and microwave sensors. Such an array - the '**digital image**' - can readily be used to drive a computer monitor or a printer after D/A conversion, this way creating an image.

Image space The mathematical space describing the (relative) positions of the observations. Image positions are expressed by their row-column index.

Imagery is mostly used as high-sounding term for images; it is avoided in this book.

Incidence angle In the context of imaging radar, the angle between the line of sight from the sensor to an element of an imaged scene and a vertical direction to the scene. One must distinguish between the nominal incidence angle determined by the geometry of the radar and the Earth's geoidal surface and the local incidence angle, which takes into account the mean slope of the ground resolution cell.

Inertial measuring unit (IMU) A device providing us with attitude data of a sensor it is attached to or of an aircraft in which it is located. An IMU is the core instrument of an INS and it is also used in a POS for direct sensor orientation.

Instantaneous field of view (IFOV) The viewing range of a detector expressed as angle, usually in milliradians; referred to in active sensing (ALS) as beam divergence of the radiation. The IFOV is one of the factors, which determine the size of the ground resolution cell of a sensor-platform system; see also **field of view**.

Instrument   In the context of GDA, a measuring device for determining the present value of a quantity under observation.

Intensity   The term is used in different disciplines with different meanings and quantifications. In this book it is used in a common language sense to express an amount of radiation, irrespective of the unit in which it is measured (radiance, irradiance, emittance, spectral emittance, etc). In radiometry, **radiant intensity** is the radiant power from a point source and measured in W sr^{-1}. See also **radiance** for comparison.

Interference   In radar RS: the wave interactions of the backscattered signals from the target surface.

Interferometry   Computational process that makes use of the interference of two coherent waves. In the case of imaging radar, two different paths for imaging cause phase differences from which an interferogram can be derived. In SAR applications, interferometry is used for constructing a DEM.

Interior orientation   See **orientation**

Interpolation   (from Latin *interpolire*, putting in between) Estimating the value of a (continuous) variable that is given by n sampled values at an intermediate point or instant. e.g. temperature has been measured at every turn of the hour; computing a temperature value at 10:37 from the 24 values given for a day is a matter of interpolation.

Interpretation elements   A set of cues used by the human vision system to interpret a picture. The seven interpretation elements are: tone/hue, texture, pattern, shape, size, height/elevation, and location/association.

Interpretation key   A guideline for image interpretation, which relates terrain features to observable features of an image.

Interpretation legend   A description of interpretation units in terms of interpretation elements.

Interval data   Data values that have some natural ordering amongst them, and that allow simple forms of arithmetic computations like addition and subtraction, but not multiplication or division. Temperature measured in centigrades is an example.

Intrinsic   A random field $Y(x)$ is said to follow the intrinsic hypothesis (or is called weak stationary) if simple differences $(Y(x) - Y(x + h))$ are stationary: there expectation exists, and their variance is independent on the location x and is solely depending on the distance h between locations.

Irradiance   The amount of incident radiation on a surface per unit area and per unit time. Irradiance is usually expressed in W m^{-2}.

Isoline   A line in the map of a spatial field that identifies all locations with the same field value. This value should be used as tag of the line, or should be derivable from tags of other lines.

Kernel   The n by m array of coefficients used by a linear filter (moving average) to compute an output pixel value from the given values at the pixel under consideration and its neighbours. A kernel can only define a linear

filter if the input (and the output) of filtering are raster data. The reciprocal of the sum of the kernel values is the 'gain' of the kernel.

Kriging   A collection of interpolation methods based on minimizing the variance of the error variance. Kriging takes into account the spatial dependence between the observations.

Laser   Acronym for Light Amplification by Stimulated Emission of Radiation. Laser instruments used as active sensors for ranging and imaging operate in the visible to near-IR range. Topographic **LIDAR** (LIght Detection And Ranging) instruments use wavelengths between 0.9 to 1.6 μm.

Latent image   When exposed to light, the silver halide crystals within the photographic emulsion undergo a chemical reaction, which results in an invisible latent image. The latent image is transformed into a visible image by the development process in which the exposed silver halide is converted into silver grains that appear black.

Latitude/Longitude (Lat/Lon)   The coordinate components of a spherical coordinate system, referred to as **geographic coordinates**. The latitude is zero on the equator and increases towards the two poles to a maximum absolute value of 90°. The longitude is counted from the Greenwich meridian positively eastwards to the maximum of 180°.

Layover   Extreme form of foreshortening, i.e., relief distortion in a radar image, in which the top of the reflecting object (e.g. a mountain) is closer to the radar than the lower part of the object. The image of such a feature appears to have fallen over towards the radar.

Least squares adjustment   A method of correcting observations in which the sum of squares of all the residuals derived by fitting the observations to a mathematical model is made a minimum. Least squares adjustment is based on probability theory and requires a (large) number of redundant measurements.

Light   Electromagnetic radiation that is visible to the human eye; also referred to as radiation enabling visual perception. We see light in the form of colour. The wavelength range is 0.38 to 0.76 μm.

Look angle   The angle of viewing relative to the vertical (nadir) as perceived from the sensor.

Map   A simplified (purpose-specific) graphical representation of geographic phenomena, usually on a planar display. From an EO perspective, it is a conventionalized image of reality, with documented conventions on the type of abstraction, the symbolic presentation, and the mathematical projection of 3D reality (including scale). Similar to stretching the notion of image to digital image we also use the term digital map.

Map coordinate system   A system of expressing the position of a point on the Earth's surface by plane rectangular coordinates using a particular map projection, such as UTM, the Lambert's conical projection, or an azimuthal stereographic projection (as used in the Netherlands).

Map generalization   The meaningful reduction of map content to accommodate scale decrease.

Mapping In the context of GDA, the process of converting (RS) data or images to a conventionalized image, the map. Mapping can be merely a radiometric and/or geometric transformation of an image, or involve information extraction by visual interpretation or automated classification.

Measurement An observation yielding a quantitative record.

Metadata Data that characterises other, usually large, data sets. For spatial data sets, this information may include volume, ownership, data format applied, spatial resolution, date of production, quality characteristics like accuracy and much more.

Microwaves Electromagnetic radiation in the microwave window, which ranges from 1 to 100 cm.

Mixel Acronym for *mixed pixel*. Mixel is used in the context of image classification where different spectral classes occur within the area covered by one pixel.

Model Simplification (abstraction) of reality.

Monochromatic (Derived form Greek *monochrōmatos*, of a single colour). Monochromatic radiation is radiation of narrow EM spectral range.

Monoplotting The process that enables extraction of accurate *(x,y)* coordinates from an image by correcting for image distortions, in particular relief displacement.

Multispectral scanning A remote sensing technique in which the Earth's surface is scanned and reflected radiation is recorded simultaneously in different wavelength bands.

Nadir The point/area on the ground vertically beneath the sensor at the moment of imaging a line or area of the scene, sometimes referred to as ground nadir. In the case of a camera, it is the vertical projection of the optical centre of the lens. The image of the ground nadir is referred to as **nadir point**, or sometimes as photograph nadir.

NDVI Normalized difference vegetation index: the difference between the surface reflectance in the NIR and the red band divided by the sum of the two.

Network A set of nodes connected by lines, representing the links between some real world entities. For example, roads, rivers or utilities that can be used for transportation analysis.

Noise Any unwanted or contaminating signal competing with the signal of interest.

Nominal data Data values that serve to identify or name something, but that do not allow arithmetic computations; sometimes also called categorical data when the values are sorted according to some set of non-overlapping categories.

Nugget effect An apparent discontinuity in the experimental variogram near the origin.

Object   An entity obtained by abstracting the real world, having a physical nature (certain composition of material), being given a descriptive name, and observable; e.g. "house". An object is a self-contained part of a scene having certain discriminating properties.

Object space   The three-dimensional region that encompasses the physical features imaged by a remote sensor; see also **image space**.

Observation   An act of recognizing and noting a fact or occurrence often involving measurement with instruments (Earth observation) as well as a record or description so obtained. The outcome can be qualitative or quantitative.

Optics   The branch of physics that deals with the propagation of light and the interaction of light with matter. While originally restricted to light, RS has stretched 'optics' to include the description of behaviour of UV and infrared radiation because of its similarity to 'sight'. Optics explains behaviour such as reflection and refraction. **Optical instruments/devices** use components such as lenses, glass prisms, and mirrors.

Orbit   The path followed by one body (e.g. a satellite) in its revolution about another (e.g. the Earth).

Ordinal data   Data values that serve to identify or name something, and for which some natural ordering of the values exists. No arithmetic is possible on these data values.

Orientation   In photogrammetry, the process of relating some form of image coordinate system to some form of 3D system. We distinguish interior, exterior, relative and absolute orientation. Interior orientation provides internal camera model information that describes the construction of the bundle of rays (as needed for exterior orientation) using the principal distance, principal point (and lens distortion). Exterior orientation provides external camera model information that describes the exact position and rotation of each image in an object space system as they existed when the image was taken. Relative orientation computes the mutual pose of two (stereoscopic) images at the time of exposure; subsequent absolute orientation establishes the relationship with an object space coordinate system.

Orthophoto/orthoimage   An aerial photo or satellite image that has been transformed to a map projection and radiometrically enhanced such that it is free of (significant) geometric distortions and radiometric disturbances.

Overlap (forward)   Considering a traditional aerial camera, when two images overlap, they share a common area. e.g. in a block or strip of photographs, adjacent/subsequent images typically overlap by 60%.

Overlay   Combining two input data layers to produce a third layer.

Panchromatic   'Sensitive to light of all colours'. For instance, the rods of our eyes are panchromatic sensors: they are not sensitive to light of a specific colour but only distinguish intensity differences of light (brightness sensed over the entire visible range). Related terms are polychromatic (multicoloured) and monochromatic (single coloured).

Parallax The apparent displacement of a point (e.g. as seen in a central perspective image) with respect to a point of reference or coordinate system, caused by the shift in the point of observation. The stereoscopic parallax - considering a stereo pair of photographs of equal principal distance - is the algebraic difference of the distances of the two images of the point from the respective nadir points of the photos measured parallel to the air base.

Pattern As an interpretation element, it refers to the spatial arrangement of features in an image; it implies the characteristic repetition of certain forms or relationships.

Pattern recognition Term for the collection of techniques used to detect and identify patterns. Patterns can be found in the spatial, spectral and temporal domains. An example of spectral pattern recognition is image classification; an example of spatial pattern recognition is segmentation.

Period The time interval of two successive maxima of an oscillation; the duration of one cycle of a wave.

Phase In the context of EM wave theory, the shift parameter for the starting point of a wave.

Photo block A block of aerial photographs is the set of images obtained by an aerial survey mission. A traditional frame camera block might consist of a number of parallel strips with a sidelap of 20- 30% and a forward overlap of 60%. Photogrammetric software typically stores all of the information associated with a photogrammetric mapping project in the **block file**, such as: spheroid and datum, map projection; camera model, ID of images, GCPs, orientation coefficients.

Photogrammetry The science and technique of making measurements on photographs and converting these to quantities meaningful in the terrain. In **analog photogrammetry**, optical and mechanical instruments, such as analog plotters, are used to reconstruct 3D geometry from two overlapping photographs. In **analytical photogrammetry** the computer replaces some expensive optical and mechanical components and software relates image space(s) to object space. In **digital photogrammetry** (hardcopy) photographs are replaced by digital images.

Photograph An image on photographic material, film or paper. A photograph in its strict sense is analogue and a record of reflected electromagnetic radiation of an object or scene of only a very narrow spectral range from ultraviolet to near infrared. In an even stricter sense, light sensitive film should also have been used by the sensor to detect the EM radiation. (Note, according to this very strict definition a picture taken by a digital camera and printed on an ink-jet printer is not a photograph nor is a panchromatic image produced from raw data of an electronic camera on a photographic film writer). **Aerial photographs** are photographs taken from positions above the Earth by a camera on an aircraft.

Photography The process or art of producing images by light on a sensitive surface (and subsequent development the exposed film).

Photon The elementary particle used to explain electromagnetic radiation in particle theory; it travels at the speed of light and has both particle and wave properties.

Picture A (2D) counterpart of an object or scene produced by a device or a human (artist). A photograph is an image, an image is a picture, but not all pictures are images and not all images are photographs.

Pixel The term stands for 'picture element'; it is the building cell of a digital image; see also **imaging sensor**.

Pixel value The digital number (DN) a pixel takes.

Platform A vehicle, such as a satellite or aircraft (or part of it), used to carry a sensor.

Polarization In the context of this book, the action or state of affecting radiation and especially EM radiation so that the vibrations of the wave assume a chosen orientation. In radar RS, we may use a signal where the electric field oscillates vertically or a microwave with a horizontally oscillation electric field, thus obtaining different images. For stereoscopic vision, we can use horizontally respectively vertically polarized light to distinguish the left image from the right image when overlaid on a computer monitor and viewed with special spectacles (one glass allowing vertically polarized light to pass through, while the other allows only horizontally polarized light).

Polychromatic 'Comprising many colours'. Solar radiation is polychromatic; see also **monochromatic**.

Polygon A computer representation of a geographic object that is perceived as a two-dimensional, i.e. area entity. The polygon is determined by a closed line that describes its boundary. Because a line is a piece-wise straight entity, a polygon is only a finite approximation of the actual area.

Prediction The prediction expresses the expected value of $Y(x)$ in a location x_0 given the measurements $x_1, x_2, \ldots, x_n$ on $X_1, X_2, \ldots, X_n$.

Principal distance The perpendicular distance from the perspective centre to the plane of an image of a digital camera or a particular finished photograph. The distance is equal to the calibrated focal length (but in the case of a film camera, corrected for the film or paper shrinkage or enlargement).

Principal point The intersection point of the perpendicular from the projection centre of a camera with the image plane.

Projective transformation A 2D geometric transformation that uses eight parameters to mathematically model perspective imaging of a plane. It defines the relationship between two coordinate systems such as the coordinate system of a central perspective image and a map coordinate system or a local coordinate system of a plane in object space; see also **conformal** and **affine transformation**.

Pulse An EM wave with a distribution confined to a short interval of time. Such a distribution is described in the time domain by its width and its amplitude.

Quantization   The number of discrete levels applied to store the energy as measured by a sensor, e.g. 8 bits quantization allows 256 levels of energy.

Radar   Acronym for Radio Detection And Ranging. Radar instruments are active sensors, sensing at wavelengths between 1 and 100 cm.

Radar equation   Mathematical expression that describes the average received signal level compared to the additive noise level in terms of system parameters. Principal parameters include the transmitted power, antenna gain, radar cross section, wavelength and range.

Radian   The SI unit of angle in a plane, equal to $180/\pi$ degrees.

Radiance   The amount of radiation being emitted or reflected per unit projected area per unit solid angle and per unit time. Radiance (observed intensity) is in photometry the radiant power from an extended area and measured in $W\ m^{-2}\ sr^{-1}$; (**power** is also called **flux** - it is radiation per unit time, measured in watt). **Spectral radiance** is radiance per wavelength (band).

Radiation   The energy of electromagnetic waves; the term radiation is most commonly used in the fields of radiometry, heating and lighting.

Radiometer   A sensor, which measures radiation and typically in one broad spectral band ('single-band radiometer') or in only a few bands ('multi-band radiometer'), but with high radiometric resolution.

Radiometric resolution   The degree to which intensity levels of incident radiation are differentiated by a sensor; usually expressed as the number of bits used for storing a DN. The number of bits used defines the quantization increment in A/D conversion. In one bit recording a radiance becomes either 0 or 1. We can show DNs on a computer monitor as grey values. If the raw data represent a binary image, then 0 is displayed as black and 1 as white. In the case of 8 bits recording we can distinguish 256 grey values for generating an image on a monitor. High-end digital cameras record in 16 bits, while older computer monitors can only support 8 bits per pixel.

Range   (a) Distance. Radar RS distinguishes 'near range', 'far range', 'slant range', 'swath range', and 'ground range'.
(b) Interval.

Range (of variogram)   The range of a variogram is the maximum distance separating points of a regionalized variable that has any significant statistical dependence. The range is the smallest variogram argument for which the variogram is either exactly equal to the sill or asymptotically close to the sill.

Raster   A set of regularly spaced (and contiguous) 2D cells with associated (field) values. In contrast to a grid, the associated values represent "cell values", not "point values". This means that the value for a cell is assumed to be valid for all locations within the cell.

Ratio data   Data values that allow most, if not all, forms of arithmetic computation, including multiplication, division, and interpolation. Typically used for cell values in raster representations of continuous fields.

Ray   (a) A beam of radiation (e.g. light). (b) Geometrically: any of a group of lines diverging from a common centre.

Reference plane   In a topocentric coordinate system, the tangential plane to the Earth ellipsoid at the nadir of the image (thus defining a **3D rectangular terrain coordinate system**, the X-axis usually oriented eastward, the Y-axis northward, and the Z-axis upward perpendicular to the reference plane).

Reflectance   The proportion of the incident radiation on a surface that is reflected; it is usually expressed as a percentage. We call **spectral reflectance** the reflectance as a function of wavelength. Reflectance is sometimes also expressed as ratio with a value range 0 to 1 and then occasionally called reflectivity.

Reflection   happens when a light ray meets a reflecting surface, such as a mirror. When light emitted by, e.g. the Sun is directed onto the surface of a mirror at a certain angle with respect to the surface normal (called incidence angle), it will be redirected into space at an angle, which is equal to the incidence angle (called the reflected angle).

Reflectivity   is not a measure of the ability of reflective learning but a measure used in telecommunication and radar to quantify a target's reflection efficiency.

Refraction   Light travels in a straight line unless it hits another medium. When light passes from one medium to another (e.g. from air to water, or air to glass), it makes a deviation from its original straight path. This phenomenon is called refraction. Since the amount of deviation depends on the wavelength, we can see the effect of refraction in the form of a rainbow, or the splitting of white light into coloured light of the rainbow-spectrum as it passes through a glass prism.

Regionalized variable   A regionalized variable is a single-value function defined over a metric space. It is used to describe natural phenomena that are characterized by fluctuations which are smooth at a global scale but erratic at a local scale. Geostatistics models regionalized variable theory as realizations of random functions.

Relief displacement   Elevation dependent shift of an imaged object by a camera. The magnitude and direction of the shift does not only depend on elevation but also on the position of the object in the image, and the camera platform characteristics.

Remote sensing (RS)   The art, science, and technology of observing an object, scene, or phenomenon by instrument-based techniques. 'Remote' because observation is done at a distance without physical contact with the object of interest.

Replicability (of image interpretation)   refers to the degree of correspondence of image interpretation results obtained by different persons for the same area or by the same person for the same area at different instants.

Resampling   The process to generate a raster with another orientation or a different cell size than the original digital image and to assign DN-values using one of the following methods: nearest neighbour selection, bilinear interpolation, or bicubic interpolation.

Residual   The difference between any measured quantity and the adjusted /computed value for that quantity.

Rods   Photoreceptor cells in the retina of the eye for sensing brightness; see also **cones**.

Roughness   A term with different meanings, among them the variation of surface elevation within a ground resolution cell. A surface appears rough to microwave illumination when the elevation variations become larger than a fraction of the radar wavelength.

Sampling   (a) Selecting a representative part of a population for statistical analysis; to this end various strategies can be applied, such as random sampling, systematic sampling, stratified sampling, etc. In signal processing, (b) 'sampling' is the process of turning a continuous signal into a discrete one to achieve A/D conversion.

Satellite   A manufactured object or vehicle intended to orbit the Earth, the moon, or another celestial body.

Scale   A word of many different meanings. The map scale or photo scale is the ratio of a distance in the image and the corresponding horizontal distance in the terrain. The ratio is commonly expressed as 1:m, where m is the **scale factor**. e.g. 1:25,000.

Scanner   (a) 'Optical scanner', not a radio receiver scanning frequencies, nor a medial scanner, nor a desktop or photogrammetric scanner): an electro-optical remote sensor with only one or a few detectors and a scanning device. The most widely used scanning device is a moving mirror. Most scanners are multispectral scanners. A laser scanner is also an optical scanner, but an active sensor emitting and sensing monochromatic EM radiation. (b) An office scanner or a photogrammetric scanner converts a hardcopy document, map, or photo to a digital image; these are typically 'flatbed scanners' using an entire linear CCD array as scanning device, which moves over the document that is mounted on a glass plate.

Scene   Section of space and time in the real world.

Segmentation   The process of dividing into segments. In **image segmentation** we aim at delineating regions that are homogeneous with respect to chosen spatial or radiometric characteristics. We can, e.g. obtain a binary image (where each region consists exclusively of pixels with either the value 0 or the value 1) by thresholding the DNs, or we want to identify image segments such that each one is homogeneous according to a criterion on texture, etc.

Sensor   In the context of this book, an instrument that detects and records EM radiation. An **active sensor** is a device that generates itself the radiation it senses. A **passive sensor** detects radiation of an external source (solar, or terrestrial, or atmospheric radiation).

Sill   The limiting value for large arguments of a variogram.

Slant range   Distance as measured by the radar to each reflecting point in the scene and recorded in the side-looking radar image.

Sliver   Small artifact polygon produced during a overlay procedure that results from the difference in accuracy between the input layers.

Spatial autocorrelation   The principle that locations which are closer together are more likely to have similar values than locations that are far apart. Often referred to as *Tobler's first law of Geography*.

Spatial data   In the broad sense, spatial data is any data with which position is associated.

Spatial Data Infrastructure   (SDI); The relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data.

Spatial database   A database that allows users to store, query and manipulate collections of georeferenced data.

Spatial interpolation   Any technique that allows to infer some unknown property value of a spatial phenomenon from values for the same property of nearby spatial phenomena. The underlying principle is that nearby things are most likely rather similar. Many spatial interpolation techniques exist.

Spatial resolution   The degree to which an image can differentiate spatial variation of terrain features. Sometimes it is specified in the image space as pixel size, or lines per millimetre (lp/mm) for photographs. More relevant for applications is the specification in object space as ground sampling distance, or ground resolution cell size as determined by the IFOV.

Speckle   Interference of backscattered waves stored in the cells of a radar image. It causes the return signals to be extinguished or amplified resulting in random dark and bright pixels in the image.

Spectral band   The interval of the EM spectrum to which the detector of a sensor is sensitive. The detector averages the spectral radiances within this range. A 'broadband sensor' such as the panchromatic camera of WorldView-1 averages per pixel the spectral response in the wavelength range from 0.4 to 0.9 μm. The spectral band of some SWIR cameras is 0.8 to 2.5 μm, while hyperspectral sensors have many but very narrow spectral bands.

Spectral reflectance curve   The curve showing the portion of the incident radiation that is reflected by a material as a function of wavelength. Sometimes called *spectral signature*.

Spectral resolution   The degree to which the spectral response of a sensor is differentiated; specified as spectral band width.

Spectral response curve   The curve portraying the sensitivity of a detector (e.g. a CCD) to radiation per wavelength. The spectral sensitivity is a detector property and should not be confused with spectral reflectance curve, which portrays a terrain property.

Spectrometer   A sensor, which measures radiance typically in many, narrow, contiguous spectral bands, usually in the visible to SWIR range of the spectrum; it offers a high spectral resolution, but low radiometric resolution as compared to a radiometer. Imaging spectrometers produce "hyperspectral images".

Specular reflection Mirror-like reflection; "bounced-off radiation" as opposed to diffuse reflection.

Static map Fixed map (e.g. a paper map, possibly scanned for dissemination through the World Wide Web) of which the contents and/or their cartographic representation cannot be changed by the user.

Steradian (symbol sr) The SI unit of solid angle. It is used to describe two-dimensional angular spans in three-dimensional space, analogously to radian in a plane. 2π sr corresponds to a hemisphere.

Stereo Short for stereoscopic. Stereoscopic viewing gives a three-dimensional impression. **Stereoscopy** is the science of producing three-dimensional visual models using two-dimensional images. We can make use of stereoscopy to make 3D measurements of objects.

Stereo model A 3D relief model observed through stereoscopic vision of a stereo pair.

Stereo pair A pair of overlapping photos or images that (partially) cover the same area from a different position. When appropriately taken, stereo pairs form a stereo model.

Stereograph A stereo pair arranged such (on the computer monitor, or on the table, or in a device) that you can readily get a 3D visual impression. Also referred to as 'stereogram'.

Stereoplotting The process that allows to measure accurate (x, y, z) coordinates from stereo models.

Stereoscopic vision The ability to perceive distance or depth by observation with both eyes. In remote sensing, stereoscopic vision is used for the three-dimensional observation of two images (photos) that are taken from different positions. Stereoscopy is used in visual image interpretation and stereoscopic measurements (stereoplotting).

Stratified sampling Taking an equal amount of samples per strata; see also **sampling**.

Subtractive colours The subtractive principle of colours is based on the three printing colours: cyan, magenta and yellow. All printed colours can be produced by a combination of these three colours. The subtractive principle is also used in colour photography.

Sun-synchronous Specification of a satellite orbit that is designed in such a way that the satellite always passes the same location on the Earth at the same local time.

Synthetic aperture radar (SAR) The (high) azimuth resolution (direction of the flight line) is achieved through off-line processing. The SAR is able to function as if it has a large virtual antenna aperture, synthesized from many observations with the (relative) small real antenna of the SAR system.

Terrain Terrain relief + terrain features (encompassing land and water).

Terrain coordinate system A system of expressing the position of a point in object space. Popular in EO are 3D rectangular coordinate systems (e.g. a topocentric system; see also **reference plane**) and hybrid horizontal-vertical systems, where horizontal position is either defined by Lat-Lon or by 2D rectangular map coordinates.

Terrain elevation Elevation of a terrain point. A **terrain point** can be a point on the ground, or on a tree, a building, a water surface. Examples of terrain elevation data are the data sets stemming from SRTM or SPOT-5 HRS.

Terrain features Land cover, all kind of topographic objects that coincide with the ground surface or 'stick out' (the roads, buildings, trees, water bodies, etc), and any other characteristics of terrain except terrain relief.

Terrain relief Ground surface: its shape, not its composition nor its cover.

Terrain surface Envelop surface of terrain relief and terrain features (as represented by a DSM).

Texture A visual surface property; the word stems from weaving. Texture as an interpretation element expresses the spatial arrangement of tonal differences in an image.

Thiessen polygons A partitioning of the plane using a given set of points and resulting in a set of polygons. Each polygon contains just one point and is the area defined by those locations that are closest to this point, and not another point in the input set. There is a natural correspondence with the Delaunay triangulation obtained from the same points.

Tone Among others one of the interpretation elements: the relative brightness in a black-and-white image.

Topography The description of a place where we live and move around.

Topological consistency The set of rules that determines what are valid spatial arrangements of simplicial complices in a spatial data representation. A typical rule is for instance that each 1-simplex must be bounded by two 0-simplices, which are its end nodes.

Topology Topology refers to the spatial relationships between geographical elements in a data set that do not change under a continuous transformation.

Transmission The process of passing on radiation through a medium or material.

Transmittance The amount of transmitted radiation by the material under consideration, expressed either as percentage of transmitted to incident radiation, or sometimes as ratio with a value range 0 to 1 and then occasionally called **transmissivity**.

Trend surface A 2D curved surface that is fitted through a number of point measurements, as an approximation of the continuous field that is measured.

Triangulated Irregular Network (TIN); a data structure that allows to represent a continuous spatial field through a finite set of (*location*, *value*) pairs and triangles made from them. Commonly in use as digital terrain model, but can be used for geographic fields other than elevation.

Tuple   A record or row in a database table; it will have several attribute values.

Variable, interval   A variable that is measured on a continuous scale. An interval variable is similar to an ordinal variable, except that the intervals between the values of the interval variable are equally spaced, so the difference between two values is meaningful. Different from a ratio variable it has no clear definition of zero, so it cannot be used to form ratios. e.g. temperature measured in Celsius (0 °C is not 'no temperature').

Variable, nominal   A variable that is organized in classes, with no natural order, i.e., cannot be ranked. It is also called a categorical variable.

Variable, ordinal   A variable that is organized in classes with a natural order, and so it can be ranked.

Variable, ratio   A variable that is measured on a continuous scale and with a clear definition of zero, so can be used to form ratios. e.g. temperature measured in Kelvin (0 K is 'no temperature').

Variance   The variance is a measure of the dispersion around the mean.

Variogram   A fucntion describing degree of spatial dependence of a random variable.

Variogram (experimental)   An estimate of the variogram based on sampling.

Vector format   Computer representation that explicitly stores the georeference of every feature.

Vegetation index   Mathematical transformation applied on surface reflectance in several spectral bands aiming at highlighting a vegetation property. There are many vegetation indices described in literature, among which NDVI is best known.

Visual variable   (Also: graphic variable); an elementary way in which graphic symbols are distinguished from each other. Commonly, the following six visual variables are recognized: *size, (lightness) value, texture, colour, orientation* and *shape*.

Volume scattering   The process of multiple reflections and redirection of radiation caused by heterogeneous material (atmosphere, water, vegetation cover, etc).

Wavelength   The distance between two successive maxima of a periodic wave (in space); mostly stated in μm or nm.

Web portal   World wide web based entrance to a spatial data clearinghouse.

Index

The core of GIScience: a process-based approach

Author
ITC

This book presents an integrated approach towards the principles of GI Science. It is a fully reworked and revised edition of earlier versions of the books Principles of Remote Sensing and Principles of Geographic Information Systems that as a two volume set have been used for teaching in ITC during many years. This one single volume presents both Earth observation by means of Remote Sensing, and GIS as a system to integrate different layers of information, including Spatial Data Infrastructure, databases and modeling. New chapters have been included on the system Earth, on integration of Remote Sensing and GIS layers and on users of GI Science with an emphasis on developing countries. In order to achieve this, a new conceptual approach to GI Science is presented in the opening chapter. This concept is fully related to current scientific developments in the field, where the integration of GIS and Remote Sensing is rapidly progressing. A key feature is that the integrating concept is based on geophysical and social processes as they occur at the earth surface. The new concept identifies the different components contained in the book. The volume is illustrated with a range of applications from various application domains and from a large diversity of countries all over the world.

Features
· Focuses on processes as occurring at the Earth surface that are in principle observable from remote sensing and earth observation technology
· Provides up-to-date remote sensing and earth observation sensors and methodology
· Provides recent, solid GeoInformation stirage and processing methodology
· Presents an introduction into data integration
· Includes a description of modern use and users of GI Science
· Is illustrated with a range of motivating and modern case studies

The present volume balances theory and applications, technology and models, observation and processing, storage and visualization, the use and the user, the process and its observations. It serves as an introduction to a large range of practical and further theoretical studies.

ITC Educational textbook series

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION