PRINCIPLES OF

GEOGRAPHIC INFORMATION SYSTEMS

ITC EDUCATIONAL TEXTBOOK SERIES *1*

Principles of Geographic Information Systems

An introductory textbook

Editor

Rolf A. de By

Authors

Rolf A. de By
Martin C. Ellis
Yola Georgiadou
Wolfgang Kainz
Richard A. Knippers
Menno-Jan Kraak
Mostafa M. Radwan
Edmund J. Sides
Yuxian Sun
Michael J. C. Weir
Cees J. van Westen

Cover illustration of the printed book:
Paul Klee (1879–1940), *Chosen Site* (1927)
Pen-drawing and water-colour on paper. Original size: 57.8×40.5 cm.
Private collection, Munich

Cover page design: Wim Feringa

Published by:
The International Institute for Aerospace Survey and Earth Sciences (ITC),
Hengelosestraat 99,
P.O. Box 6,
7500 AA Enschede, The Netherlands

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Principles of Geographic Information Systems
Rolf A. de By (ed.)
(ITC Educational Textbook Series; 1)
Second edition
ISBN 90–6164–200-0 ITC, Enschede, The Netherlands
ISSN 1567–5777 ITC Educational Textbook Series

Contents

List of Figures

List of Tables

Preface

This book was designed for a three-week lecturing module on the principles of geographic information systems, to be taught to students in all education programmes at ITC as the second module in their course.

A geographic information system is a computer-based system that allows to study natural and man-made phenomena with an explicit bearing in space. To this end, the GIS allows to enter data, manipulate the data, and produce interpretable output that may teach us lessons about the phenomena.

There are many uses for GIS technology, and ITC, with all its different domains of scientific applications, is the living proof of this statement. Fields we have in mind are, for instance, soil science, management of agricultural, forest and water resources, urban planning, geology, mineral exploration, cadastre and environmental monitoring. It is likely that the student reader of this textbook is already a domain expert in one of these fields; the intention of the book is to lay the foundation to also become an expert user of GIS technology.

With so many different fields of application, it is impossible to single out the specific techniques of GIS usage for all of them in a single book. Rather, the book focuses on a number of common and important topics that any GIS user should be aware of to be called an expert user. We further believe that GIS is going to

be used differently in the future, and that our students should now be provided with a broad foundation, so as to be effective in their use of GIS technology then as well.

The book is also meant to define a common understanding and terminology for follow-up modules, which the student may elect later in the programme. The textbook does not stands by itself, but was developed in synchrony with the textbook on *Principles of Remote Sensing* [30].

Structure of the book

The chapters of the book have been arranged in a rather classical set-up. Chapter 1 to 3 provide a generic introduction to the field, discussing the geographic phenomena that interest us (Chapter 1), the ways these phenomena can be represented in a computer system (Chapter 2), and the data processing systems that are used to this end (Chapter 3).

Chapter 4 to 6 subsequently focus on the *process* of using a GIS environment. We discuss how spatial data can be obtained, entered and prepared for use (Chapter 4), how data can be manipulated to improve our understanding of the phenomena that they represent (Chapter 5), and how the results of such manipulations can be visualized (Chapter 6). Special attention throughout these chapters is devoted to the specific characteristics of *spatial* data. In the last chapter, we direct our attention to the issue of the quality of data and data manipulations, as a lesson of what we can and cannot read in GIS output (Chapter 7).

Each chapter contains sections, a summary and some exercises. The exercises are meant to be a test of understanding of the chapter's contents; they are not practical exercises. They may not be typical exam questions either!

Besides the regular chapters, the back part of the book contains a bibliography, a glossary, an index, and an appendix that lists a number of important internet sites.

Electronic version of the book

The book is also made available as an electronic document, with hyperlinks to pages, references, figures, tables and websites. The purpose of this electronic version is twofold: it can be used as an on-line aid in studying the material; in the future, it allows the authors to use the document as a 'coat rack' to add answers to existing questions, add new questions (and their answers), provide errata to the original text, new websites and other information that may become available. The electronic version of the book can be browsed but not be printed.

How to read the book

This book is the intended study material of a three week module, but it is not the only material to help the student master the topics covered. In each education programme, lectures and practicals have been developed to also aid in bringing the knowledge across. The best advice for the student is to read the book in synchrony with the lectures offered during the module. This will ease the understanding and allow to timely pose the questions that may arise.

For some students, some chapters or parts of chapters will be easier to study single-handedly than for others. Differences in professional and training backgrounds are more prominent in ITC's student population than possibly anywhere else. It is important to understand one's strengths and weaknesses and to take appropriate action by seeking help where needed. The book contains important material as it provides a foundation of a number of other teaching modules, later in the curriculum.

Throughout, a number of textual conventions have been applied, most of them in line with [34], [41] and [59]. Chapters are arranged in sections, and these possibly in subsections. The table of contents provides an overview. Important terms are *italicized*, and many of these can be found in the index, some of them even in the glossary.

Not all the text in this book is compulsory study material for all students! Sections with a caution traffic sign in the margin, as the one found on the left, indicate that this part of the book is optional.[1] The lecturer will indicate whether these parts must be studied in your programme.

[1] The idea of such a signpost comes from [34].

Acknowledgements

The book has already quite a history, with a predecessor for the 1999 curriculum. This is a heavily revized, in parts completely rewritten, version of that first edition. Much of the work for the first edition, besides that of the authors and editors—including Cees van Westen—was in the capable hands of Erica Weijer, supported by Marion van Rinsum. Ineke ten Dam supervised the whole production process in 1999 as well as in the year 2000.

Many people were instrumental to the production of the current book, first and foremost, obviously, the authors of respective (parts of) chapters. Their names are found on the title sheet. Kees Bronsveld and Rob Lemmens were the careful and critical readers of much of the text, and provided valuable suggestions for improvements. Connie Blok, Allan Brown, Corné van Elzakker, Yola Georgiadou, Lucas Janssen, Barend Köbben and Bart Krol read and commented upon specific chapters. Rob Lemmens and Richard Knippers provided additional exercises. Jan Hendrikse provided help in the mathematics of digital elevation models.

Many illustrations in the book come from the original authors, but have been restyled for this publication. The technical advice of Wim Feringa in this has been crucial, as has his work on the cover plate. A number of illustrations was produced from data sources provided by Sherif Amer, Wietske Bijker, Wim Feringa, Robert Hack, Asli Harmanli, Gerard Reinink, Richard Sliuzas, Siefko Slob, and Yuxian Sun. In some cases, because of the data's history, they can perhaps be better ascribed to an ITC division: Cartography, Engineering Geology, and Urban Planning and Management.

Finally, this book would not have materialized in its present form without the dedication of and pleasant collaboration with Lucas Janssen, the editor of

the sister textbook to this volume, *Principles of Remote Sensing*.

Technical account

This book was written using Leslie Lamport's LaTeX generic typesetting system, which uses Donald Knuth's TeX as its formatting engine. Figures came from various sources, but many were eventually prepared with Macromedia's Freehand package, and then turned into PDF format.

From the LaTeX sources we generated the book in PDF format, using the PDFLaTeX macro package, supported by various add-on packages, the most important being Sebastian Rahtz' hyperref.

Rolf A. de By, Enschede, September 2000

Preface to the second edition

This second edition of the GIS book is an update of the first edition, with many (smaller) errors removed. I am grateful to all the students who pointed out little mistakes and inconsistencies, or parts in the text that were difficult to understand. A special word of thanks goes to Wim Bakker, for his, at points almost annoying, meticulous proofreading and keen eye for finer detail. A number of colleagues made valuable comments that helped me work on improving the text as well.

Many parts of the text have remained fundamentally unchanged. Improvements, I believe, have been made on the issue of spatiotemporal data models in Chapter 2. The section on three-dimensional data analysis in Chapter 5 has been taken out, as it was no longer felt to be 'core material'. The discussion of error propagation in Chapter 7 has also been elaborated upon substantially.

A book like this one will never be perfect, and the field of GIS has not yet reached the type of maturity where debates over definitions and descriptions are no longer needed. As always, I will happily receive comments and criticisms, in a continued effort to improve the materials.

Rolf A. de By, Enschede, September 2001

Chapter 1

A gentle introduction to GIS

1.1 The purpose of GIS

Students from all over the world visit ITC to attend courses. They often stay for half a year, but many of them stay longer, perhaps up to 18 months. Some eventually find a position as Ph.D. student—usually after successfully finishing a regular M.Sc. course. If we attempt to define what is the common factor in the interests of all these people, we might say that they are involved in studies of their environment, in the hope of a better understanding of that environment. By *environment*, we mean the geographic space of their study area and the events that take place there.

For instance

- an *urban planner* might like to find out about the urban fringe growth in her/his city, and quantify the population growth that some suburbs are witnessing. S/he might also like to understand why it is *these* suburbs and not others;
- a *biologist* might be interested in the impact of slash-and-burn practices on the populations of amphibian species in the forests of a mountain range to obtain a better understanding of the involved long-term threats to those populations;
- a *natural hazard analyst* might like to identify the high-risk areas of annual monsoon-related flooding by looking at rainfall patterns and terrain characteristics;
- a *geological engineer* might want to identify the best localities for constructing buildings in an area with regular earthquakes by looking at rock formation characteristics;

- a *mining engineer* could be interested in determining which prospect copper mines are best fit for future exploration, taking into account parameters such as extent, depth and quality of the ore body, amongst others;
- a *geoinformatics engineer* hired by a telecommunication company may want to determine the best sites for the company's relay stations, taking into account various cost factors such as land prices, undulation of the terrain *et cetera*;
- a *forest manager* might want to optimize timber production using data on soil and current tree stand distributions, in the presence of a number of operational constraints, such as the requirement to preserve tree diversity;
- a *hydrological engineer* might want to study a number of water quality parameters of different sites in a freshwater lake to improve her/his understanding of the current distribution of *Typha* reed beds, and why it differs so much from that of a decade ago.

All the above professionals work with data that relates to space, typically involving positional data. Positional data determines *where* things are, or perhaps, where they were or will be. More precisely, these professionals deal with questions related to *geographic space*, which we might informally characterize as having positional data relative to the Earth's surface.

Positional data of a non-geographic nature is not of our interest in this book. A car driver might want to know where is the head light switch; a surgeon must know where is the appendix to be removed; NASA must know where to send its spaceships to Mars. All of this involves positional information, but to use the Earth's surface as a reference for these purposes is not a good idea.

The acronym GIS stands for *geographic information system*. A GIS is a computerized system that helps in maintaining data about geographic space. This is its primary purpose. We provide a more elaborate definition in Section 1.1.2. But first, let us try to make some clear observations about our points of departure.

1.1.1 Some fundamental observations

Our world is constantly changing, and not all changes are for the better. Some changes seem to have natural causes (volcano eruptions, meteorite impacts) while others are caused by man (for instance, land use changes or land reclamation from the sea, a favourite pastime of the Dutch). There is also a large number of global changes for which the cause is unclear: think of the greenhouse effect and global warming, the El Niño/La Niña events, or, at smaller scales, landslides and soil erosion.

For background information on El Niño, take a look at Figure 1.1. It presents information related to a study area (the equatorial Pacific Ocean), with positional data taking a prominent role. We will use the study of El Niño as an example of using GIS for the rest of this chapter.

In summary, we can say that changes to the Earth's geography can have *natural* or *man-made* causes, or a *mix of both*. If it is a mix of causes, we usually do not quite understand the changes fully.

We, humans, are an inquisitive breed. We want to understand what is going on in our world, and this is why we study the phenomena of geographic change. In many cases, we want to deepen our understanding, so that there will be no more unpleasant surprises; so that we can take action when we feel that action must be taken. For instance, if we understand El Niño better, and can forecast that another event will be in the year 2004, we can devise an action plan to reduce the expected losses in the fishing industry, to lower the risks of landslides caused by heavy rains or to build up water supplies in areas of expected droughts.

The fundamental problem that we face in many uses of GIS is that of understanding phenomena that have (a) a *geographic dimension*, as well as (b) a *temporal dimension*. We are facing 'spatio-temporal' problems. This means that our object of study has different characteristics for different locations (the geographic di-

El Niño is an aberrant pattern in weather and sea water temperature that occurs with some frequency (every 4–9 nine years) in the Pacific Ocean along the Equator. It is characterized by less strong western winds across the ocean, less upwelling of cold, nutrient-rich, deep-sea water near the South American coast, and therefore by substantially higher sea surface temperatures (see figures below). It is generally believed that El Niño has a considerable impact on global weather systems, and that it is the main cause for droughts in Wallacea and Australia, as well as for excessive rains in Peru and the southern U.S.A.
El Niño means 'little boy' because it manifests itself usually around Christmas. There exists also another—less pronounced–pattern of *colder* temperatures, that is known as La Niña. La Niña occurs less frequently than El Niño. The figures below left illustrate an extreme El Niño year (1997; considered to be the most extreme of the twentieth century) and a subsequent La Niña year (1998).
Left figures are from December 1997, and extreme El Niño event; right figures are of the subsequent year, indicating a La Niña event. In all figures, colour is used to indicate sea water temperature, while arrow lengths indicate wind speeds. The top figures provide information about absolute values, the bottom figures about values relative to the average situation for the month of December. The bottom figures also give an indication of wind speed and direction. See also Figure 1.3 for an indication of the area covered by the array of buoys.
At the moment of writing, August 2001, another El Niño event, not so extreme as the 1997 event, is forecasted to occur at the end of the year 2001.

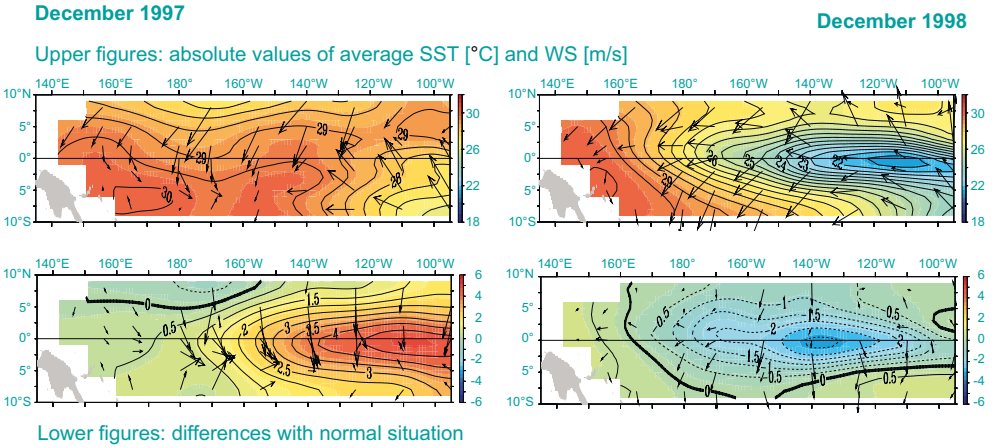


Figure 1.1: The El Niño event of 1997 compared with a more normal year 1998. The top figures indicate average Sea Surface Temperature (SST, in colour) and average Wind Speed (WS, in arrows) for the month of December. The bottom figures illustrate the anomalies (differences from a normal situation) in both SST and WS. The island in the lower left corner is (Papua) New Guinea with the Bismarck Archipelago. Latitude has been scaled by a factor two. Data source: National Oceanic and Atmospheric Administration, Pacific Marine Environmental Laboratory, Tropical Atmosphere Ocean project (NOAA/PMEL/TAO).

mension) and that it has different characteristics for different moments in time (the temporal dimension).

The El Niño event is a good example of such a phenomenon, because (a) sea surface temperatures differ between locations, and (b) sea surface temperatures change from one week to the next.

1.1.2 A first definition of GIS

Let us take a closer look at the El Niño example. Many professionals study that phenomenon closely, most notably meteorologists and oceanographers. They prepare all sorts of products, such as the maps of Figure 1.1, to improve their understanding. To do so, they need to obtain data about the phenomenon, which obviously here will include measurements about sea water temperature and wind speed in many locations. Next, they must process the data to enable its analysis, and allow interpretation. This interpretation will benefit if the processed data is presented in an easy to interpret way.

We may distinguish three important stages of working with geographic data:

Data preparation and entry The early stage in which data about the study phenomenon is collected and prepared to be entered into the system.

Data analysis The middle stage in which collected data is carefully reviewed, and, for instance, attempts are made to discover patterns.

Data presentation The final stage in which the results of earlier analysis are presented in an appropriate way.

We have numbered the three phases, and thereby indicated the most natural order in which they take place. But such an order is only a sketch of an ideal situation, and more often we find that a first attempt of data analysis suggests that we need more data. It may also be that the data representation leads to follow-up questions for which we need to do more analysis, for which we may be needing more data. This shows that the three phases may be iterated over a number of times before we are happy with our work. We look into the three phases more below, in the context of the El Niño project.

Data preparation and entry

In the El Niño case, our data acquisition means that the project collects sea water temperatures and wind speed measurements. This is achieved by mooring buoys with measuring equipment in the ocean. Each buoy measures a number of things: wind speed and direction, air temperature and humidity, sea water temperature at the surface and at various depths down to 500 metres. Our discussion focuses on sea surface temperature (SST) and wind speed (WS). A typical buoy is illustrated in Figure 1.2, which shows the placement of various sensors on the buoy.

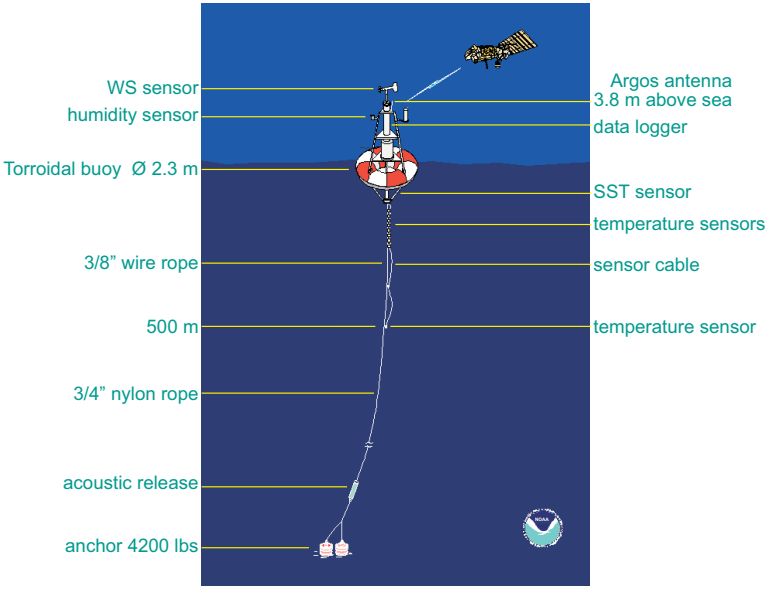


Figure 1.2: Schematic overview of an ATLAS type buoy for monitoring sea water temperatures in the El Niño project

For monitoring purposes, some 70 buoys were deployed at strategic places

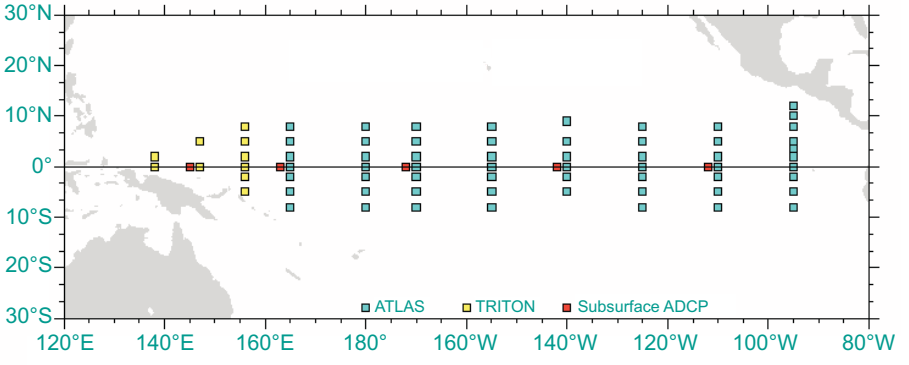


Figure 1.3: The array of positions of sea surface temperature and wind speed measuring buoys in the equatorial Pacific Ocean

within 10° of the Equator, between the Galapagos Islands and New Guinea. Figure 1.3 provides a map that illustrates the positions of these buoys. The buoys have been anchored, so they are stationary. Occasional malfunctioning is caused sometimes by high seas and bad weather or by getting entangled in long-line fishing nets. As Figure 1.3 shows, there happen to be three types of buoy, but we will not discuss their differences.

All the data that a buoy obtains through thermometers and other sensors with which it is equipped, as well as the buoy's geographic position is transmitted by satellite communication daily. This data is stored in a computer system. We will from here on assume that acquired data has been put in digital form, that is, it has been converted into computer-readable format.

In the textbook on *Principles of Remote Sensing* [30], many other ways of acquiring geographic data will be discussed. During the current module, we will assume the data has been obtained and we can start to work with it.

Data analysis

Once the data has been collected in a computer system, we can start analysing it. Here, let us look at what processes were probably involved in the eventual production of the maps of Figure 1.1.[1] Observe that the production of maps belongs to the phase of data presentation that we discuss below.

Here, we look at how data generated at the buoys was processed before map production. A closer look at Figure 1.1 reveals that the data being presented are based on the monthly averages for SST and WS (for two months), not on single measurements for a specific date. Moreover, the two lower figures provide comparisons with 'the normal situation', which probably means that a comparison was made with the December averages for a long series of years.

Another process performed on the initial (buoy) data is that they have been generalized from 70 point measurements (one for each buoy) to cover the complete study area. Clearly, for positions in the study area for which no data was available, some type of interpolation took place, probably using data of nearby buoys. This is a typical GIS function: deriving the value of a property for some location where we have not measured.

It seems likely that the following steps took place for the upper two figures. We look at SST computations only—WS analysis will have been similarly conducted:

1. For each buoy, using the daily SST measurements for the month, the average SST for that month was computed. This is a simple computation.

2. For each buoy, the monthly average SST was taken together with the geo-

[1] We say 'probably' because we are not participating in the project, and we can only make an educated guess at how the data was actually operated upon.

Buoy	Geographic position	Dec. 1997 avg. SST
B0789	(165° E, 5° N)	28.02 °C
B7504	(180° E, 0° N)	27.34 °C
B1882	(110° W, 7°30′ S)	25.28 °C
...	...	...

Table 1.1: The georeferenced list (in part) of average sea surface temperatures obtained for the month December 1997.

graphic location, to obtain a *georeferenced* list of averages, as illustrated in Table 1.1.

3. From this georeferenced list, through a method of *spatial interpolation*, the estimated SST of other positions in the study are were computed. This step was performed as often as needed, to obtain a fine mesh of positions with measured or estimated SSTs from which the maps of Figure 1.1 were eventually derived.

4. We assume that previously to the above steps we had obtained data about average SST for the month of December for a long series of years. This too may have been spatially interpolated to obtain a 'normal situation' December data set of a fine granularity.

Let us clarify what is meant by a 'georeferenced' list first. Data is *georeferenced* (or spatially referenced) if it is associated with some position using a spatial reference system. This can be by using (longitude, latitude) coordinates, or by other means that we come to speak of in Chapter 4. The important thing is to have an agreed upon coordinate system as a reference. In our list, we have associated average sea surface temperatures with positions, and thereby we have georeferenced them.

In step 3 above, we mentioned spatial interpolation. To understand this issue, first observe that sea surface temperature is a property that occurs everywhere in the ocean, and not only at buoys. The buoys only provide a finite sample of the property of sea surface temperature. *Spatial interpolation* is a technique that allows us to estimate the value of a property (SST in our case) also in places where we have not measured it. To do so, it uses measurements of nearby buoys.[2]

The theory of spatial interpolation is extensive, but this is not the place to discuss it. It is however a typical example of data manipulation that a GIS can perform on user data.

[2]There are in fact many different spatial interpolation techniques, not just one.

Data presentation

After the data manipulations discussed above, our data is prepared for producing the maps of Figure 1.1. The data representation phase deals with putting all together into a format that communicates the result of data analysis in the best possible way.

Many issues come up when we want to have an optimal presentation. We must consider what is the message we want to bring across, who is the audience, what is the presentation medium, which rules of aesthetics apply, and what techniques are available for representation. This may sound a little abstract, so let us clarify with the El Niño case.

For Figure 1.1, we made the following observations:

- The *message* we wanted to bring across is to illustrate what are the El Niño and La Niña events, both in absolute figures and in relative figures, i.e., as differences from a normal situation.
- The *audience* for this data presentation clearly were the readers of this text book, i.e., students of ITC who want to obtain a better understanding of GIS.
- The *medium* was this book, so, printed matter of A4 size, and possibly also a website. The book's typesetting imposes certain restrictions, like maximum size, font style and font size.
- The *rules of aesthetics* demanded many things: the maps should be printed with north up, west left; with clear georeference; with intuitive use of symbols et cetera. We actually also violated some rules of aesthetics, for instance, by applying a different scaling factor in latitude compared to longitude.

- The *techniques* that we used included use of a colour scheme, use of isolines,[3] some of which were tagged with their temperature value, plus a number of other techniques.

[3] Isolines are discussed in Chapter 2.

GIS defined

So, what is a GIS? In a nutshell, we can define a *geographic information system* as a computerized system that facilitates the phases of data entry, data analysis and data presentation especially in cases when we are dealing with georeferenced data.

This means that a GIS user will expect support from the system to enter (georeferenced) data, to analyse it in various ways, and to produce presentations (maps and other) from the data. Many kinds of functionality should come with this: support for various kinds of coordinate systems and transformations between them, many different ways of 'computing' with the georeferenced data, and obviously a large degree of freedom of choice in presentation parameters such as colour scheme, symbol set, medium et cetera.

We will later make the subtle distinction between a *GIS* and a *GIS application*. For now it suffices to give an example of this often missed subtlety. We discussed above a GIS application: determining sea water temperatures of the El Niño event in two subsequent December months. The same software package that we used to do this analysis could tomorrow be used to analyse forest plots in northern Thailand, for instance. That would mean another GIS application, but using the same GIS. Hence, a GIS is the software package that can (generically) be applied to many different applications. When there is no risk of ambiguity, people sometimes do not make the distinction between a 'GIS' and a 'GIS application'.

1.1.3 Spatial data and geoinformation

Another subtle difference exists between the terms *data* and *information*. Most of the time, we use the two terms almost interchangeably, and without the risk of being ambiguous. Occasionally, however, we need to be precise and then their distinction matters.

By *data*, we mean representations that can be operated upon by a computer. More specifically, by *spatial data* we mean data that contains positional values. Occasionally one will find in the literature the more precise phrase *geospatial data* as a further refinement, which then means spatial data that is georeferenced. (Strictly speaking, spatial data that is not georeferenced can have positional data unrelated to the Earth's surface. Examples can be found in molecular chemistry, in which the position of atoms in molecules are defined relative to each other, and in industrial design engineering, in which the parts of a car engine are defined relative to each other.) In this book, we will use 'spatial data' as a synonym for 'georeferenced data'.

By *information*, we mean data that has been interpreted by a human being. Humans work with and act upon information, not data. Human perception and mental processing leads to information, and hopefully understanding and knowledge. One cannot expect a machine like a computer to 'understand' or 'have knowledge'. *Geoinformation* is a specific type of information that involves the interpretation of spatial data.

1.1.4 Applications of GIS

There are many different uses of GIS, as may have become clear from our list of professionals on page 27 who deal with geoinformation. Throughout this book, we will provide examples of different types of GIS use, hopefully by the end having covered a fair number of scientific areas in which ITC is active.

An important distinction between GIS applications is whether the geographic phenomena studied are *man-made* or *natural*. Clearly, setting up a cadastral information system, or using GIS for urban planning purposes involves a study of man-made things mostly: the parcels, roads, sidewalks, and at larger scale, suburbs and transportation routes are all man-made. These entities often have—or are assumed to have—clear-cut boundaries: we know, for instance, where one parcel ends and another begins.

On the other hand, geomorphologists, ecologists and soil scientists often have natural phenomena as their study objects. They may be looking at rock formations, plate tectonics, distribution of natural vegetation or soil units. Often, these entities do not have clear-cut boundaries, and there exist transition zones where one vegetation type, for instance, is gradually replaced by another.

It is not uncommon, of course, to find GIS applications that do a bit of both, i.e., they involve both natural and man-made entities. Examples are common in areas where we study the effect of human activity on the environment. Railroad construction is such an area: it may involve parcels to be reclaimed by government, it deals with environmental impact assessment and will usually be influenced by many restrictions, such as not crossing seasonally flooded lands, and staying within inclination extremes in hilly terrain.

A second distinction in applications of GIS stems from the overall purposes of use of the system. A prototypical use of GIS is that of a research project with an explicitly defined project objective. Such projects usually have an *a priori* defined

duration. Feasibility studies like site suitability, but also simulation studies, for instance in erosion modelling, are examples. We call all of these *project-based* GIS applications.

In contrast to these are what we call *institutional* GIS applications. They can be characterized in various ways. The life time (duration) of these applications is either indefinite, or at least not *a priori* defined. Their goal is usually to provide base data to others, not to address a single research issue. Good examples of this category are monitoring systems like early warning systems for food/water scarcity, or systems that keep track of weather patterns. Indeed, our El Niño example is best qualified under this heading, because the SST and WS measurements continue. Another class of examples is found in governmental agencies like national topographic surveys, cadastral organizations and national census bureaus. They see it as their task to administer (geographic) changes, and their main business is to stay up-to-date, and provide data to others, either (more historically) in the form of printed material such as maps or (more recently) in the form of digital data.

1.2 The real world and representations of it

When dealing with data and information we usually are trying to represent some part of the real world as it is, as it was, or perhaps as we think it will be. A computerized system can help to store such representations. We restrict ourselves to 'some part' of the real world simply because it cannot be represented completely. The question which part must be represented should be entirely answered through the notion of *relevance* to the purpose of the computerized system.

The El Niño system discussed earlier in this chapter has as its purpose the administration of SST and WS in various places in the equatorial Pacific Ocean, and to generate georeferenced, monthly overviews from these. If this is its complete purpose, the system should not store data about the ships that moored the buoys, the manufacture date of the buoys *et cetera*. All this data is irrelevant for the purpose of the system.

The fact that we represent the real world only in part teaches us to be humble about the expectations that we can have about the system: all the data it can possibly generate for us in the future must in some way be made available to it first.

In general, a computer representation of some part of the real world, if set-up in a good way, will allow us to enter and store data, analyse the data and transfer it to humans or to other systems. We will now look at setting up real world representations.

1.2.1 Modelling

'Modelling' is a buzzword, used in many different ways and many different meanings. A representation of some part of the real world can be considered a *model* of that part. We call it such because the representation will have certain characteristics in common with the real world. This allows us to study the representation, i.e., the model, instead of the real world. The advantage of this is that we can 'play around' with the model and look at different scenarios, for instance, to answer 'what if' questions. We can change the data in the model, and see what are the effects of the changes.

Models—as representations—come in many different flavours. In the GIS environment, the most familiar model is that of a *map*. A map is a miniature representation of some part of the real world. Paper maps are the best known, but digital maps also exist, as we shall see in Chapter 6. We look more closely at maps below.

Another important class of models are databases. A database stores a usually considerable amount of data, and provides various functions to operate on the stored data. Obviously, we will be especially interested in databases that store spatial data.

The phrase 'data modelling' is the common name for the design effort of structuring a database. This process involves the identification of the kinds of data that the database will store, as well as the relationships between these data kinds. In data modelling, the most important tool is the *data model*, and we come back to it in Section 1.2.3. 'Spatial data modelling' is a specific type of data modelling that we will also discuss there.

Maps and databases can be considered *static models*. At any point in time, they represent a single state of affairs. Usually, developments or changes in the real world are not easily recognized in these models. *Dynamic models* or *process*

models address precisely this issue. They emphasize changes that have taken place, are taking place or may take place. Dynamic models are inherently more complicated than static models, and usually require much more computation to obtain an intuitive presentation of the underlying processes. Simulation models are an important class of dynamic models that allow to simulate real world processes.

Observe that our El Niño system can be called a static model as it stores state-of-affairs data such as the average December 1997 temperatures. But at the same time, it can also be considered a simple dynamic model, because it allows us to compare different states of affairs, as Figure 1.1 demonstrates. This is perhaps the simplest dynamic model: a series of 'static snapshots' allows us to infer some information about the behaviour of the system.

1.2.2 Maps

The best known (conventional) models of the real world are maps. Maps have been used for thousands of years to represent information about the real world. Their conception and design has developed into a science with a high degree of sophistication. Maps have proven to be extremely useful for many applications in various domains.

A disadvantage of maps is that they are restricted to two-dimensional static representations, and that they always are displayed in a given scale. The map scale determines the spatial resolution of the graphic feature representation. The smaller the scale, the less detail a map can show. The accuracy of the base data, on the other hand, puts limits to the scale in which a map can be sensibly drawn. The selection of a proper map scale is one of the first and most important steps in map design.

A map is always a graphic representation at a certain level of detail, which is determined by the scale. Map sheets have physical boundaries, and features spanning two map sheets have to be cut into pieces.

Cartography as the science and art of map making functions as an interpreter translating real world phenomena (primary data) into correct, clear and understandable representations for our use. Maps also become a data source for other maps.

With the advent of computer systems, analogue cartography became digital cartography. It is important to note that whenever we speak about cartography today, we implicitly assume digital cartography. The use of computers in map making is an integral part of modern cartography. The role of the map changed accordingly. Increasingly, maps lose their role as data storage. This role is taken over by (spatial) databases. What remains is the visualization function of maps.

1.2.3 Databases

A *database* is a repository capable of storing large amounts of data. It comes with a number of useful functions:

1. the database can be used by multiple users at the same time—i.e., it allows *concurrent use*,
2. the database offers a number of techniques for storing data and allows to use the most efficient one—i.e., it supports *storage optimization*,
3. the database allows to impose rules on the stored data, which will be automatically checked after each update to the data—i.e., it supports *data integrity*,
4. the database offers an easy to use data manipulation language, which allows to perform all sorts of data extraction and data updates—i.e., it has a *query facility*,
5. the database will try to execute each query in the data manipulation language in the most efficient way—i.e., it offers *query optimization*.

Databases can store almost any sort of data. Modern database systems, as we shall see in Section 3.3, organize the stored data in tabular format, not unlike that of Table 1.1. A database may have many such tables, each of which stores data of a certain kind. It is not uncommon that a table has many thousands of data rows, sometimes even hundreds of thousands.

For the El Niño project, one may assume that the buoys report their measurements on a daily basis and that these measurements are stored in a single, large table.

DAYMEASUREMENTS

Buoy	*Date*	*SST*	*WS*	*Humid*	*Temp10*	...
B0749	1997/12/03	28.2 °C	NNW 4.2	72%	22.2 °C	...
B9204	1997/12/03	26.5 °C	NW 4.6	63%	20.8 °C	...
B1686	1997/12/03	27.8 °C	NNW 3.8	78%	22.8 °C	...
B0988	1997/12/03	27.4 °C	N 1.6	82%	23.8 °C	...
B3821	1997/12/03	27.5 °C	W 3.2	51%	20.8 °C	...
B6202	1997/12/03	26.5 °C	SW 4.3	67%	20.5 °C	...
B1536	1997/12/03	27.7 °C	SSW 4.8	58%	21.4 °C	...
B0138	1997/12/03	26.2 °C	W 1.9	62%	21.8 °C	...
B6823	1997/12/03	23.2 °C	S 3.6	61%	22.2 °C	...
...	...	...	...	...	...	...

Table 1.2: A stored table (in part) of daily buoy measurements. Illustrated are only measurements for December 3rd, 1997, though measurements for other dates are in the table as well. *Humid* is the air humidity just above the sea, *Temp10* is the measured water temperature at 10 metres depth. Other measurements are not shown.

The El Niño buoy measurements database likely has more tables than the one illustrated. There may be data available about the buoys' maintenance and service schedules; there may be data about the gauging of the sensors on the buoys, possibly including expected error levels. There will almost certainly be a table that stores the geographic location of each buoy.

Table 1.1 was obtained from table DAYMEASUREMENTS through the use of the data manipulation language. A query was defined that computes the monthly average SST from the daily measurements, for each buoy. A discussion of the data manipulation language that was used is beyond the purpose of this book, but we should mention that the query was a simple, four-line program.

A database design determines which tables will be present and what sort of columns (attributes) each table will have. A completed database design is known as the *database schema*. To define the database schema, we use a language,

commonly known as a *data model*. Confusingly perhaps, a data model is not a model in the sense of what we discussed before. It is not a model of any kind, but rather a language that can be used to define models. It is the use of this language, and hence the definition of a model that we call data modelling, and which results in a database schema.

1.2.4 Spatial databases

Spatial databases are a specific type of database. They store representations of geographic phenomena in the real world to be used in a GIS. They are special in the sense that they use other techniques than tables to store these representations. This is because it is not easy to represent geographic phenomena using tables. We will not discuss these more appropriate techniques in this book.

A spatial database is not the same thing as a GIS, though they have a number of common characteristics. A spatial database focuses on the functions we listed above for databases in general: concurrency, storage, integrity, and querying, especially, but not only, spatial data. A GIS, on the other hand, focuses on operating on spatial data with what we might call a 'deeper understanding' of geographic space. It knows about spatial reference systems, and functionality like distance and area computations, spatial interpolations, digital elevation models *et cetera*. Obviously, a GIS must also store its data, and for this it provided relatively rudimentary facilities. More and more, we see GIS applications that use the GIS for the spatial analysis, and a separate spatial database for the data storage.

The assumption for the design of a spatial database schema is that the relevant spatial phenomena exist in a two- or three-dimensional Euclidean space. Euclidean space can be informally defined as a model of space in which locations are represented as coordinates—(x, y) in 2D; (x, y, z) in 3D—and notions like *distance* and *direction* have been defined, with the usual formulas. In 2D, we also talk about the *Euclidean plane*.

The phenomena that we want to store representations for in a spatial database may have point, line, area or image characteristics. Different storage techniques exist for each of them. An important choice in the design of a spatial database application is whether some geographic phenomenon is better repre-

sented as a point, as a line or as an area. Currently, the support for image data exists but is not impressive. Some GIS applications may even be more demanding and require point representations in certain cases, and area representation in other cases. Cities on a map may have to be represented as points or as areas, depending on the scale of the map.

To support this, the database must store representations of geographic phenomena (spatial features) in a scaleless and seamless manner. *Scaleless* means that all coordinates are world coordinates given in units that are normally used to reference features in the real world (using a spatial reference system). From such values, calculations can be easily performed and any (useful) scale can be chosen for visualization. A *seamless* database does not show map sheet boundaries or other partitions of the geographic space other than imposed by the spatial features themselves. This may seem a trivial remark, but early GIS applications had map production as their prime purpose, and considered map sheet boundaries as important spatial features.

All geographic phenomena have various relationships among each other and possess spatial (geometric), thematic and temporal attributes (they exist in space and time). Phenomena are classified into thematic data layers depending on the purpose of the database. This is usually described by a qualification of the database as, for example, a cadastral, topographic, land use, or soil database. A spatial database not only serves to store the data and manipulate it, as it should also allow the users to carry out simple forms of spatial analysis.

Spatial analysis involves questions about the data that relate topological and other relationships. Such questions may involve neighbourhood, distance, direction, incidence, disjointness and a few more characteristics that may exist among geographic phenomena. In the El Niño case, for example, we may want to find out where is epicentre of warm water or where is the steepest gradient in water

temperature.

GIS and databases

A database, like a GIS, is a software package capable of storing and manipulating data. This begs the question when to use which, or possibly when to use both. Historically, these systems have different strengths, and the distinction remains until this day.

Databases are good at storing large quantities of data, they can deal with multiple users at the same time, they support data integrity and system crash recovery, and they have a high-level, easy to use data manipulation language. GISs are not very good at any of this.

GIS, however, is tailored to operate on *spatial* data, and allows all sorts of analysis that are inherently geographic in nature. This is probably GIS's main stronghold: combining in various ways the representations of geographic phenomena. GIS packages, moreover, nowadays have wonderful, highly flexible tools for map production, of the paper and the digital type. GIS have an embedded 'understanding' of geographic space. Databases mostly lack this type of understanding.

The two, however, are growing towards each other. All good GIS packages allow to store the base data in a database, and to extract it from there when needed for GIS operation. This can be achieved with some simple settings and/or program statements inside the GIS. Databases, likewise, have moved towards GIS and many of them nowadays allow to store spatial data also in different ways. Previously, they in principle were capable of storing such data, but the techniques were fairly inefficient.

In summary, one might conclude that small research projects can probably be carried out without the use of a real database. GIS have rudimentary database facilities on board; the user should be aware they are really rudimentary. Mid-sized projects use a database/GIS tandem for data storage and manipula-

tion. Larger projects, long-term projects and institutional projects organize their spatial data processing around a spatial database, not around a GIS. They use the GIS mostly for spatial analysis and output presentation. We will look more closely at these data processing systems in Chapter 3.

1.3 An overview of upcoming chapters

In this chapter, we provided an introduction to the area of GIS, by means of an example study of the El Niño phenomenon.

In Chapter 2, we will focus the discussion on different kinds of geographic phenomena and their representation in spatial data, and try to build more intuition for these different phenomena and data, also in terms of when to use which.

Chapter 3 is devoted to a much more in-depth study of the two data processing systems for spatial data, namely, GIS and databases. We will discuss backgrounds and typical types of use of these systems.

The Chapter 4 and 5 focus on actual use of a GIS. The first especially looks at the phase of data entry and preparation: how to ensure that the (spatial) data is correctly entered into the GIS, such that it can be used in subsequent analysis. The most important forms of spatial data analysis are discussed in the latter chapter.

The phase of data presentation, also known as *visualization*, is the topic of Chapter 6. It involves a discussion of cartographic principles: what to put on a map, where to put it, and what techniques to use. Sooner or later, each serious GIS user will be involved also in output presentation (usually of maps), so it is important to understand the underlying principles.

The final Chapter 7 addresses the rather general and important issues of spatial data quality. As will become clear in discussions of the Principles of Remote Sensing, spatial data never has infinite precision, and usually has some error. Errors may cause certain manipulations to become meaningless, so awareness of the GIS user on this subject is important.

Summary

This chapter provides an elementary discussion of what is GIS. Technical details have been mostly left out, as building some sound intuition was the main purpose.

We looked at the purposes of GIS and identified understanding our geographic space as the main thread amongst all GIS applications. We saw that spatial data and spatial data processing are important factors in that understanding, and that GIS are built to do this. A simple example of a study of the EL Niño effect provided an illustration, although again we skipped the technical details.

The use of GIS commonly takes place in three phases: data entry, data analysis and data presentation.

Representations are models of real world phenomena. In areas close to geography, we saw that maps have been in use for a long time. More recently, databases were used as digital models of real world phenomena. GIS are specifically created to define geographic models of the real world.

Digital models (as in a database or GIS) have enormous advantages over paper models (such as maps). They are more flexible, and therefore much easier changed for the purposes at hand. They, in principle, allow automatic animations and simulations, carried out by the computer system on which the software runs. This has opened up an important toolbox that may help to improve our understanding of the world.

The attentive reader will have noted our threefold use of the word 'model'. This, perhaps, may be confusing. Except as a verb, where is means 'to describe' or 'to represent', it is also used as a noun. A 'real world model' is a representation of a number of phenomena in the real world, usually to enable some type of administration, computation and/or simulation. It is the result of the activity

of 'modelling'. A 'data model', on the other hand, is a database language used as a tool in database design, in an activity called 'data modelling'. The result of that activity is not a data model, but a database schema, an abstract definition of the contents of the (future) database. A database schema can be viewed as a special kind of real world model, but it is abstract because it identifies only types of things in that real world, and not the things itself. Therefore, we might say that a database schema is an occurrence-independent real world model.

Project-based GIS applications usually have a clear-cut purpose, for instance, to improve the understanding of some spatial phenomenon. These applications can be short-lived: the research is carried out by collecting data, entering in the GIS, analysing the data, and producing informative maps. An example is rapid earthquake damage assessment.

Institutional GIS applications, on the other hand, usually have as their goal the continued administration of spatial change and the sustained availability of spatial base data. Their needs for advanced data analysis are usually less, and the complexity of these applications lies more in the continued provision of trustworthy data to others. They are thus long-lived applications. An obvious example are automated cadastral systems.

Questions

1. Take another look at the list of professions provided on page 27. Give two more examples of professions that people are trained in at ITC, and describe a possible relevant problem in their 'geographic space'.

2. In Section 1.1.1, some examples are given of changes to the Earth's geography. They were categorized in three types: *natural changes*, *man-made changes* and *somewhere-in-between*. Provide additional examples of each category.

3. What kind of professionals, do you think, were involved in the Tropical Atmosphere Ocean project of Figure 1.1? Hypothesize about how they obtained the data to prepare the illustrations of that figure. How do you think they came up with the nice colour maps?

4. Use arguments obtained from Figure 1.1 to explain why 1997 was an El Niño year, and why 1998 was not. Also explain why 1998 was in fact a La Niña year, and not an ordinary year.

5. On page 35, we made the observation that we would assume the data that we talk about to have been put into a digital format, so that computers can operate on them. However, much useful data has not been converted in this way. Provide examples from your own experience of data sources in non-digital format. (You may even consider the question how these sources could be made digital, but strictly speaking this is a topic we will only discuss in Chapter 4.)

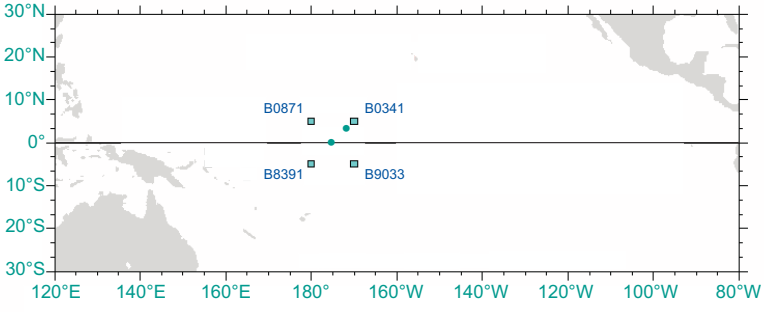


Figure 1.4: Just four measuring buoys

6. Assume the El Niño project is operating with just four buoys, and not 70, and their location is as in Figure 1.4. We have already computed the average SSTs for the month December 1997, which are provided in the table below. Answer the following questions:

   - What is the expected average SST of the illustrated location that is precisely in the middle of the four buoys?
   - What can be said about the expected SST of the illustrated location that is closer to buoy B0341? Make an educated guess at the temperature that could have been observed there.

Buoy	*Position*	*SST*
B0341	(160° W, 6° N)	30.18 °C
B0871	(180° W, 6° N)	28.34 °C
B8391	(180° W, 6° S)	25.28 °C
B9033	(160° W, 6° S)	28.12 °C

7. The categorization of GIS applications in Section 1.1.4 provides two important distinctions that are independent of each other. This leads to four types of GIS application. What are they? Give a good example of each.

8. Argue why scale is not important in spatial data storage, whether in the GIS or in a separate spatial database. Provide (exceptional) cases of applications or spatial data use, in which scale may matter in spatial data storage.

9. In Table 1.2, we illustrated some stored measurements data. The table uses one row of data for each day that a buoy reports its measurements. How many rows do you think the table will store after a full year of project execution?

   The table does *not* store the geographic location of the buoy involved. Why do you think it doesn't do that? How do you think are these locations stored?

10. On page 52, we discussed Euclidean space and the Euclidean plane. We simply mentioned that distance and direction are defined with the *usual* formulas, without mentioning them. Provide the usual formula for the distance between two locations, (x_1, y_1) and (x_2, y_2), in the Euclidean plane.

    Under what condition(s) can we say that the first location lies north of the second location? Under what condition can we say that it lies west of it?

Chapter 2

Geographic information and Spatial data types

In the previous chapter, we identified geographic phenomena as the study objects of the field of GIS. GIS supports such study because it represents phenomena digitally in a computer. GIS also allows to visualize these representations in various ways. Figure 2.1 provides a summary sketch.

Geographic phenomena exist in the real world: for true examples, one has to look outside the window. In using GIS software, we first obtain some *computer representations* of these phenomena—stored in memory, in bits and bytes—as faithfully as possible. This is where we speak of spatial data. We continue to manipulate the data with techniques usually specific to the application domain, for instance, in geology, to obtain a geological classification. This may result in additional computer representations, again stored in bits and bytes. For true examples of these representations, one would have to look into the files in which

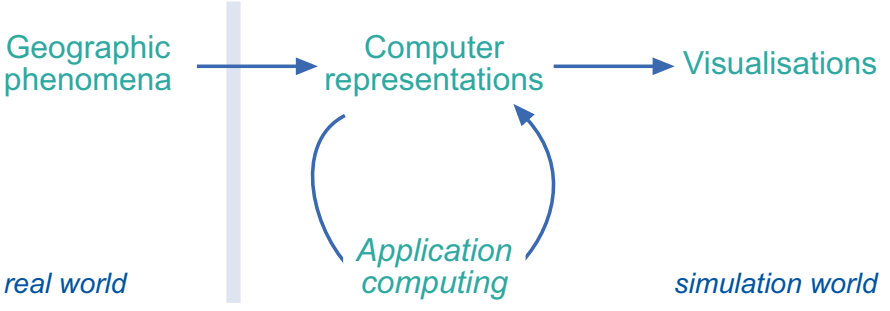


Figure 2.1: The three ways in which we can look at the objects of study in a GIS application.

they are stored. One would see the bits and bytes, but very exciting this would not be. Therefore, we can also use the GIS to create *visualizations* from the computer representation, either on-screen, printed on paper, or otherwise.

It is crucial to understand the fundamental differences between these three notions. The real world, after all, is a completely different domain than the GIS/computer world, in which we simulate the real world. Our simulations, we know for sure, will never be perfect, so some facts may not be found.

Crossing the barrier between the real world and a computer representation of it is a domain of expertise by itself. Mostly, it is done by direct observations using sensors and digitizing the sensor output for computer usage. This is the domain of *remote sensing*, the topic of *Principles of Remote Sensing* [30] in a next module. Other techniques for obtaining computer representations are more indirect: we can take a visualization result of a previous project, for instance a paper map, and re-digitize it.

This chapter studies (types of) geographic phenomena more deeply, and looks into the different types of computer representations for them. Any geographic phenomenon can be represented in various ways; the choice which representa-

tion is best depends mostly on two issues:

- what original, raw data (from sensors or otherwise) is available, and
- what sort of data manipulation does the application want to perform.

Finally, we mention that illustrations in this chapter—by nature—are visualizations themselves, although some of them are intended to illustrate a geographic phenomenon or a computer representation. This might, but should not, confuse the reader.[1] This chapter does not deal with visualizations.

[1] To this end, map-like illustrations in this chapter purposely do not have a legend or text tags. They are intended *not* to be maps.

2.1 Geographic phenomena

In the previous chapter, we discussed the reasons for taking GIS as a topic of study: they are the software packages that allow us to analyse geographic phenomena and understand them better. Now it is time to make a more prolonged excursion along these geographic phenomena and to look at how a GIS can be used to represent each of them.

There is of course a wide range of geographic phenomena as a short walk through the ITC building easily demonstrates. In the corridors, one will find poster presentations of many different uses of GIS. All of them are based on one or more notions of geographic phenomenon.

2.1.1 Geographic phenomenon defined

We might define a geographic phenomenon as something of interest that

- *can be named or described*,
- *can be georeferenced*, and
- *can be assigned a time (interval) at which it is/was present*.

What the relevant phenomena are for *one's current use* of GIS depends entirely on the objectives that one has.

For instance, in water management, the objects of study can be river basins, agro-ecologic units, measurements of actual evapotranspiration, meteorological data, ground water levels, irrigation levels, water budgets and measurements of total water use. Observe that all of these can be named/described, georeferenced and provided with a time interval at which each exists.

In multipurpose cadastral administration, the objects of study are different: houses, barns, 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.

Observe that we do not claim that all relevant phenomena come as triplets $(\mathrm{description}, \mathrm{georeference}, \mathrm{time\_interval})$, though many do. If the georeference is missing, we seem to have something of interest that is not positioned in space: an example is a legal document in a cadastral system. It is obviously somewhere, but its position in space is considered irrelevant.

If the time interval is missing, we seem to have a phenomenon of interest that is considered to be always there, i.e., the time interval is (likely to be considered) infinite. If the description is missing, ..., we have something funny that exists

in space and time, yet cannot be described. (We do not think such things can be interesting in GIS usage.)

Referring back to the El Niño example discussed in Chapter 1, one could say that there are at least three geographic phenomena of interest there. One is the Sea Surface Temperature, and another is the Wind Speed in various places. Both are phenomena that we would like to understand better. A third geographic phenomenon in that application is the array of monitoring buoys.

2.1.2 Different types of geographic phenomena

Our discussion above of what are geographic phenomena was necessarily abstract, and therefore perhaps somewhat difficult to grasp. The main reason for this is that geographic phenomena come in so many different 'flavours'. We will now try to categorize the different 'flavours' of geographic phenomena.

To this end, first make the observation that the representation of 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 skip over the time part for now, and come back to that issue in Section 2.4. The reason why we ignore temporal issues is that current GIS do not provide much automatic support for time-dependent data, and that it must be considered an issue of advanced GIS use.

A second fundamental observation is that some phenomena manifest themselves essentially everywhere in the study area, while others only occur in certain localities. If we define our study area as the equatorial Pacific Ocean, for instance, we can say that Sea Surface Temperature can be measured anywhere in the study area. Therefore, it is a typical example of a (geographic) *field*.

> A (geographic) **field** is a geographic phenomenon for which, for every point in the study area, a value can be determined.

The usual examples of geographic fields are temperature, barometric pressure and elevation. These fields are actually continuous in nature. Examples of discrete fields are land use and soil classifications. Again, any location is attributed a single land use class or soil class. We discuss fields further in Section 2.1.3.

Many other phenomena do not manifest themselves everywhere in the study area, but only in certain localities. The array of buoys of the previous chapter is

a good example: there is a fixed number of buoys, and for each we know exactly where it is located. The buoys are typical examples of (geographic) *objects*.

(Geographic) **objects** populate the study area, and are usually well-distinguishable, discrete, bounded entities. The space between them is potentially empty.

A general rule-of-thumb is that natural geographic phenomena are more often fields, and man-made phenomena are more often objects. Many exceptions to this rule actually exist, so one must be careful in applying it. We look at objects in more detail in Section 2.1.4.

Elevation in the Falset study area, Tarragona province, Spain. The area is approximately 25 $\times$ 20 km. The illustration has been aesthetically 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 continuous function over the space.

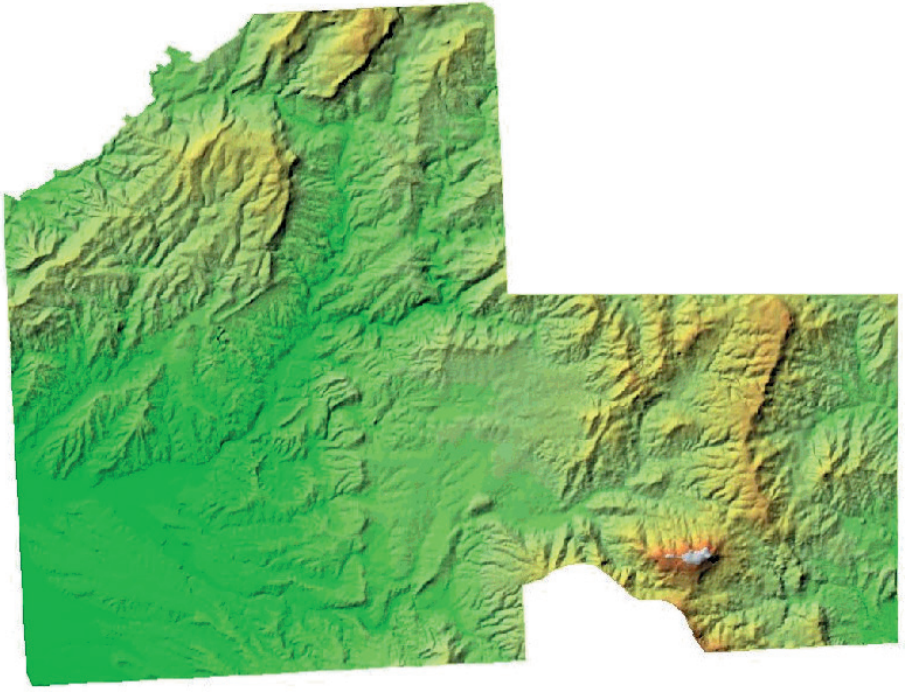


Figure 2.2: A continuous field example, namely the *elevation* in the study area. Data source: Division of Engineering Geology (ITC)

2.1.3 Geographic fields

A field is a geographic phenomenon that has a value 'everywhere' in the study space. We can therefore think of a field f as a function from any position in the study space to the domain of values of the field. If (x, y) is a position in the study area then $f(x, y)$ stands for the value of the field f at locality (x, y).

Fields can be *discrete* or *continuous*, and if they are continuous, they can even be differentiable.

In a *continuous field*, the underlying function is assumed to be continuous, such as is the case for temperature, barometric pressure or elevation. Continuity means that all changes in field values are gradual. A continuous field can even be *differentiable*. In a differentiable field we can determine a measure of change (in the field value) per unit of distance anywhere and in any direction. 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.

Figure 2.2 illustrates the variation in elevation in a study area in Spain. A colour scheme has been chosen to depict that variation. This is a typical example of a continuous field.

There are many variations of non-continuous fields, the simplest example being elevation in a study area with perfectly vertical cliffs. At the cliffs there is a sudden change in elevation values. An important class of non-continuous fields are the discrete fields. *Discrete fields* cut up 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 identifying geological units in the Falset study area—

is provided in Figure 2.3. Observe that locations on the boundary between two parts can be assigned the field value of the 'left' or 'right' part of that boundary.

One may note that discrete fields are a step from continuous fields towards geographic objects: discrete fields as well as objects make use of 'bounded' features. Observe, however, that a discrete field still assigns a value to *every* location in the study area, something that is not typical of geographic objects.

A *field-based model* consists of a finite collection of geographic fields: we may be interested in elevation, barometric pressure, mean annual rainfall, and maximum daily evapotranspiration, and thus use four different fields.

pale yellows (mostly lower left): Miocene and Quaternary
dark greens (right): Cretaceous
violets: Lias
dark orange: Bundsandstein
greys: intrusive and sedimentary areas
dark greens (left): Oligocene
pale oranges: Eocene
purples: Keuper and Muschelkalk

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 (geological unit).

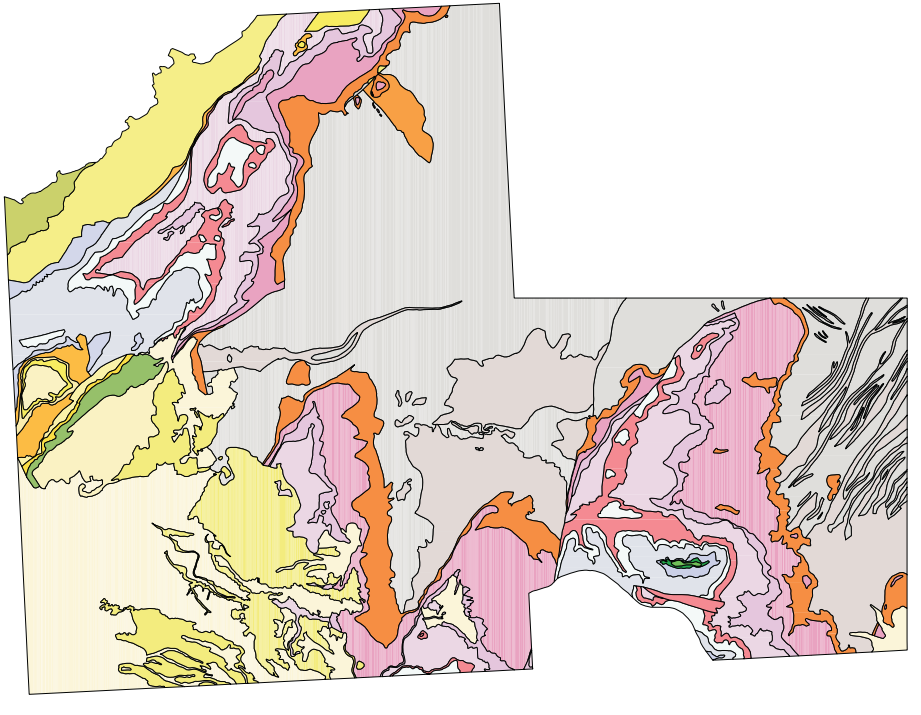


Figure 2.3: A discrete field indicating geological units, used in a foundation engineering study for constructing buildings. The same study area as in Figure 2.2. Data source: Division of Engineering Geology (ITC)

Kinds of data values

Since we have now discriminated between continuous and discrete fields, we may also look at different kinds of data values. *Nominal data values* are values that provide a name or identifier so that we can discriminate between different values, but that is about all we can do. Specifically, we cannot do true computations with these values. An example are the names of geological units. This kind of data value is sometimes also called *categorical data*.

Ordinal data values are data values that can be put in some 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.

Interval data values and *ratio data values* do allow computation. The first differs from the second in that it knows no arithmetic zero value, and does not support multiplication or division. For instance, a temperature of 20 °C is not twice as warm as 10 °C, and thus centigrade temperatures are interval data values, not ratio data values. Rational data have a natural zero value, and multiplication and division of values are sensible operators: distances measured in metres are an example.

Observe that continuous fields can be expected to have ratio data values, simply because we must be able to interpolate them.

2.1.4 Geographic objects

When the geographic phenomenon is not present everywhere in the study area, but somehow 'sparsely' populates it, we look at it in terms of *geographic objects*. Such objects are usually easily distinguished and named. Their position in space is determined by a combination of one or more of the following parameters:

- *location* (where is it?),
- *shape* (what form is it?),
- *size* (how big is it?), and
- *orientation* (in which direction is it facing?).

Several attempts have been made to define a taxonomy of geographic object types. Dimension is an important aspect of the shape parameter. It answers the question whether an object is perceived as a point feature, a linear, area or volume feature.

How we want to use the information about a geographic object determines which of the four above parameters is required to represent it. For instance, in a car navigation system, all that matters about geographic objects like petrol stations is where they are, and thus, location suffices. Shape, size and orientation seem to be irrelevant. In the same system, however, roads are important objects, and for these some notion of location (where does it 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 information components.

Shape is usually important because one of its factors is dimension: are the objects inherently considered to be zero-, one-, two- or three-dimensional? The

petrol stations mentioned above apparently are zero-dimensional, i.e., they are perceived as points in space; roads are one-dimensional, as they are considered to be lines in space. In another use of road information—for instance, in multi-purpose cadastre systems where precise location of sewers and manhole covers matters—roads might well be considered to be two-dimensional entities, i.e., areas within which a manhole cover may fall.

Figure 2.4 illustrates geological faults in the Falset study area, a typical example of a geographic phenomenon that exists of objects and that is not a field. Each of the faults has a location, and apparently for this study it is best to view a fault shaped as a one-dimensional object. The size, which is length in case of one-dimensional objects, is also indicated. Orientation does not play a role in this case.

We usually do not study geographic objects in isolation, but whole *collections of objects* viewed as a unit. These object collections may also have specific geographic characteristics.

Most of the more interesting collections of geographic objects obey certain natural laws. The most common (and obvious) of these is that different objects do not occupy the same location. This, for instance, holds for

- the collection of petrol stations in a car navigation system,
- the collection of roads in that system,
- the collection of parcels in a cadastral system,

and in many more cases. We will see in Section 2.2 that this natural law of 'mutual non-overlap' has been a guiding principle in the design of computer representations for geographic phenomena.

Observe that collections of geographic objects can be interesting phenomena at the higher aggregation level: forest plots form forests, parcels form suburbs, streams, brooks and rivers form a river drainage system, roads form a road network, SST buoys form an SST monitoring system, *et cetera*. It is sometimes useful to view the geographic phenomena also at this aggregated level and look at characteristics like coverage, connectedness, capacity and so on. Typical questions are:

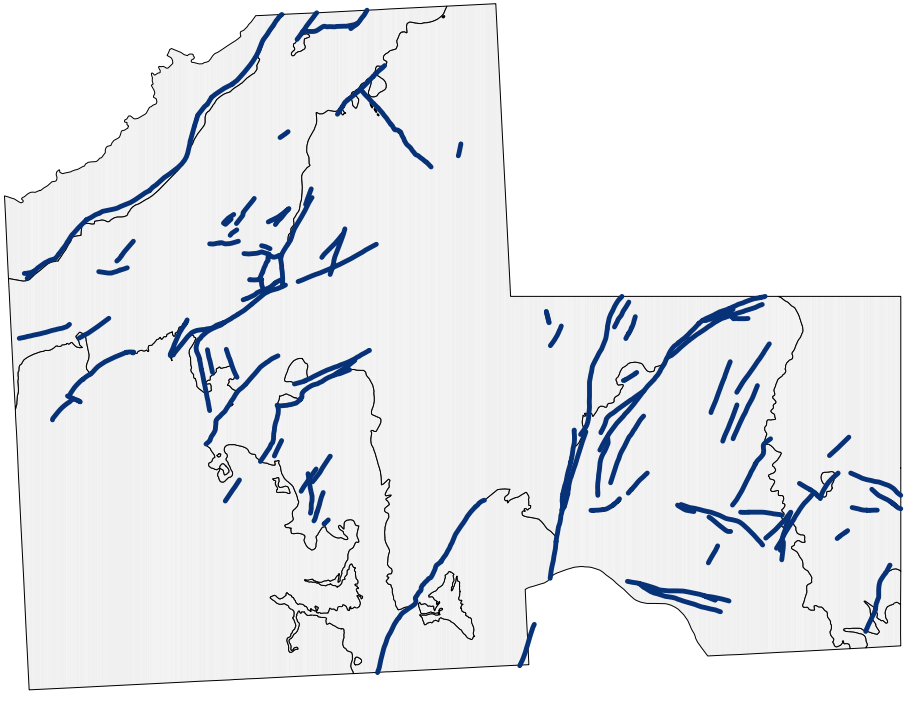


Figure 2.4: A number of geological faults in the same study area as in Figure 2.2. Faults are indicated in blue; the study area, with the main geological era's is set in grey in the background only as a reference. Data source: Division of Engineering Geology (ITC)

- Which part of the 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)

It is in this context that studies of *multi-scale* approaches are also conducted. Multi-scale approaches look at the problem of how to maintain and operate on multiple representations of the same geographic phenomenon.

Other spatial relationships between the members of a geographic object collection may exist and can be relevant in GIS usage. Many of them fall in the category of topological relationships, which is what we discuss in Section 2.2.4.

2.1.5 Boundaries

Where shape and/or size of contiguous areas matter, the notion of *boundary* comes into play. This is true for geographic objects but also for the constituents of a discrete geographic field, as will be clear from another look at Figure 2.3.

Location, shape and size are fully determined if we know an area's boundary, so the boundary is a good candidate for representing it. This is especially true for areas that have naturally *crisp* boundaries. A crisp boundary is one that can be determined with almost arbitrary precision, dependent only on the data acquisition technique applied. *Fuzzy* boundaries contrast with crisp boundaries in that the boundary is not a precise line, but rather itself an area of transition.

As a general rule-of-thumb—again—crisp boundaries are more common in man-made phenomena, whereas fuzzy boundaries are more common with natural phenomena. In recent years, various research efforts have addressed the issue of explicit treatment of fuzzy boundaries, but in day-to-day GIS use these techniques are neither often supported, nor often needed. The areas identified in a geological classification, like that of Figure 2.3, for instance, are surely vaguely bounded, but applications of this type of information probably do not require high positional accuracy of the boundaries involved, and thus, an assumption that they are actually crisp boundaries does not influence the usefulness of the data too much.

2.2 Computer representations of geographic information

Up to this point, we have not discussed at all how geoinformation, like fields and objects, is represented in a computer. One needs to understand at least a little bit about the computer representations to understand better what the system does with the data, and also what it cannot do with it.

In the above, we have seen that various geographic phenomena have the characteristics of continuous functions over geometrically bounded, yet infinite domains of space. Elevation, for instance, can be measured at arbitrarily many locations, even within one's backyard, and each location may give a different value.

When we want to represent such a phenomenon faithfully in computer memory, we could either:

- try to store as many $(\text{location}, \text{elevation})$ pairs as possible, or
- try to find a symbolic representation of the elevation function, as a formula in x and y—like $(3.0678x^2 + 20.08x - 7.34y)$ or so—which after evaluation will give us the elevation value at a given (x, y).

Both approaches have their drawbacks. The first suffers from the fact that we will never be able to store *all* elevation values for all locations; after all, there are infinitely many locations. The second approach suffers from the fact that we have no clue what such a function should be, or how to derive it, and it is likely that for larger areas it will be an extremely complicated function.

In GISs, typically a combination of both approaches is taken. We store a finite, but intelligently chosen set of locations with their elevation. This gives us

the elevation for those stored locations, but not for others. Therefore, the stored values are paired with an interpolation function that allows to infer a reasonable elevation value for locations that are not stored. The underlying principle is called *spatial autocorrelation*: locations that are close are more likely to have similar values than locations that are far apart.

The simplest interpolation function—and one that is in common use—simply takes the elevation value of the nearest location that is stored! But smarter interpolation functions, involving more than a single stored value, can be used as well, as may be understood from the SST interpolations of Figure 1.1.

Line objects, either by themselves or in their role of region object boundaries, are another common example of continuous phenomena that must be finitely represented. In real life, these objects are usually not straight, and often erratically curved. A famous paradoxical question is whether one can actually measure the length of Great Britain's coastline ... can one measure around rocks, pebbles or even grains of sand?[2] In a computer, such random, curvilinear features can never be fully represented.

One must, thus, observe that phenomena with intrinsic continuous and/or infinite characteristics have to be represented with finite means (computer memory) for computer manipulation, and that any finite representation scheme that forces a discrete look on the continuum that it represents is open to errors of interpretation.

In GIS, fields are usually implemented with a *tessellation* approach, and objects with a (topological) *vector* approach. This, however, is not a hard and fast rule, as practice sometimes demands otherwise.

[2]Making the assumption that we can decide where precisely the coastline is ... it may not be so crisp as we think.

In the following sections we discuss tessellations, vector-based representations and how these can be applied to represent geographic fields and objects.

2.2.1 Regular tessellations

A *tessellation* (or tiling) is a partition of space into mutually exclusive cells that together make up the complete study space. With each cell, some (thematic) value is associated to characterize that part of space. Three regular tessellation types are illustrated in Figure 2.5. In a regular tessellation, the cells are the same shape and size. The simplest example 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 2.5–left).

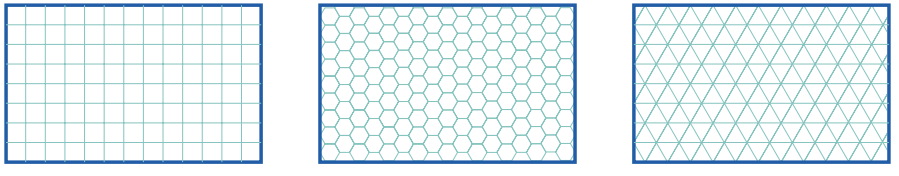


Figure 2.5: The three most common regular tessellation types: square cells, hexagonal cells, and triangular cells.

All regular tessellations have in common that the cells are of 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.

The square cell tessellation is by far the most commonly used, mainly because georeferencing a cell is so straightforward. Square, regular tessellations are known under various names in different GIS packages: *raster* or *raster map*. The size of the area that a raster cell represents is called the raster's *resolution*. Sometimes, the word *grid* is also used, but strictly speaking, a *grid* is an equally spaced collection of *points*, which all have some attribute value assigned. They are often used for discrete measurements that occur at regular intervals. Grid

points are often considered synonymous with raster cells. (See also pages 475 and 478.)

Our finite approximation of the study space leads to some forms of interpolation that must be dealt with. 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 often says that lower and left boundaries belong to the cell. To improve on this continuity issue, we can do two things:

- 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 has the continuity characteristic.

Usually, if one wants to use rasters for continuous field representation, one does the first but not the second. The second technique is usually considered too computationally costly for large rasters.

The location associated with a raster cell is fixed by convention, and may be the cell centroid (mid-point) or, for instance, its left lower corner. Values for other positions than these must be computed through some form of interpolation function, which will use one or more nearby field values to compute the value at the requested position. This allows to represent continuous, even differentiable, functions.

An important advantage of regular tessellations is that we *a priori* know how they partition space, and 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 phenomenon of interest.

Adaptivity to the phenomenon to represent can pay off. Suppose we use any of the above regular tessellations to represent elevation in a perfectly flat area. Then, clearly we need 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, although the elevation might be 1500 m above sea level everywhere.

2.2.2 Irregular tessellations

Above, we discussed that regular tessellations provide simple structures with straightforward algorithms, which are, however, not adaptive to the phenomena they represent. This is why substantial effort has also been put into *irregular tessellations*. Again, these are partitions of space into mutually disjoint cells, but now the cells may vary in size and shape, allowing them to adapt to the spatial phenomena that they represent. We discuss here only one type, namely the region quadtree, but we point out that many more structures have been proposed in the literature and have been implemented as well.

Irregular tessellations are more complex than the regular ones, but they are also more adaptive, which typically leads to a reduction in the amount of memory used to store the data.

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 together be represented as one bigger cell. A simple illustration is provided in Figure 2.6. 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, known as a quadtree. In main memory, the nodes of a quadtree (both circles and squares in the figure below) are represented as records. The links between them are pointers, a programming technique to address (i.e., to point to) other records.

Quadtrees are adaptive because they apply the *spatial autocorrelation principle*:

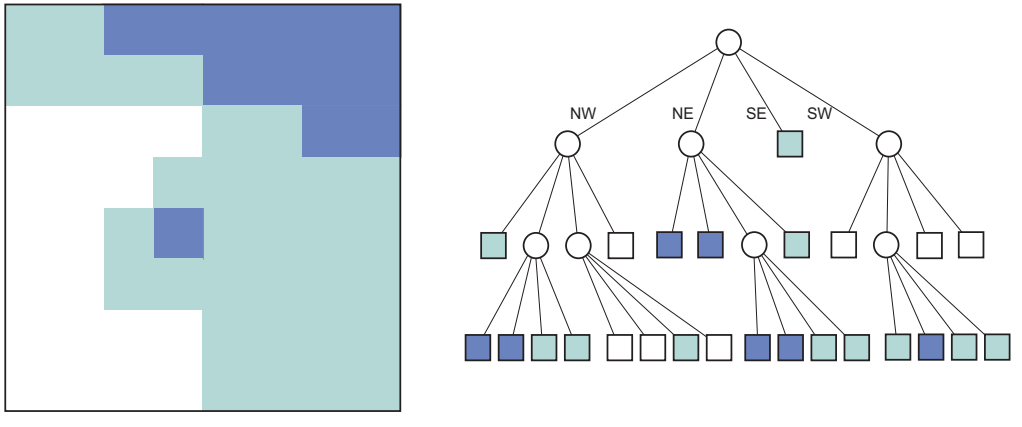


Figure 2.6: An 8×8, three-valued raster (here: colours) and its representation as a region quadtree. To construct the quadtree, the field is successively split in four quadrants until parts have only a single field value. After the first split, the southeast quadrant is entirely green, and this is indicated by a green square at level two of the tree. Other quadrants had to be split further.

locations that are near in space are likely to have similar field values. When a conglomerate of cells has the same value, they are represented together in the quadtree, provided boundaries coincide with the predefined quadrant boundaries. This is why we can also state that a quadtree provides a nested tessellation: quadrants are only split if they have two or more values (colours).

Quadtrees have various interesting characteristics. One of them is that the square nodes at the same level represent equal area sizes. This allows to quickly compute the area covered by some field value. The top node of the tree represents the complete raster.

2.2.3 Vector representations

In summary of the above, we can say that tessellations cut up the study space into cells, and assign a value to each cell. A raster is a regular tessellation with square cells, and this is by far the most commonly used. How the study space is cut up is (to some degree) arbitrary, and this means that cell boundaries usually have no bearing to the real world phenomena that are represented.

In *vector representations*, an attempt is made to associate georeferences with the geographic phenomena explicitly. A georeference is a coordinate pair from some geographic space, and is also known as a vector. This explains the name. We will see a number of examples below.

Observe that tessellations do not explicitly store georeferences of the phenomena they represent. Instead, they might 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.

Below, we discuss various vector representations. We start with our discussion with the TIN, a representation for geographic fields that can be considered a hybrid between tessellations and vector representations.

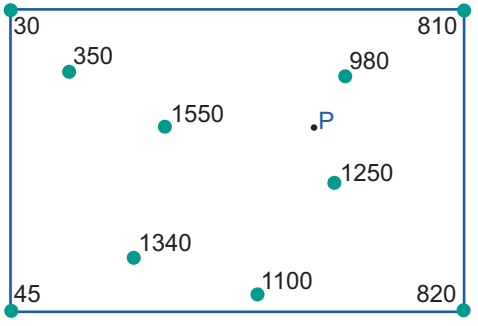


Figure 2.7: Input locations and their (elevation) values for a TIN construction. The location P is an arbitrary location that has no associated elevation measurement and that is only included for explanation purposes.

Triangulated Irregular Networks

A commonly used data structure in GIS software is the *triangulated irregular network*, or *TIN*. It is one of the standard implementation techniques for digital terrain models, but it can be used to represent any continuous field.

The principles behind a TIN 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 usually not on a nice regular grid. Any location together with its elevation value can be viewed as a point in three-dimensional space. This is illustrated in Figure 2.7. From these 3D points, we can construct an irregular tessellation made of triangles. Two such tessellations are illustrated in Figure 2.8.

Observe that 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.[3]

[3]The *slope* in a location is usually defined to consist of two parts: the *gradient* and the *aspect*.

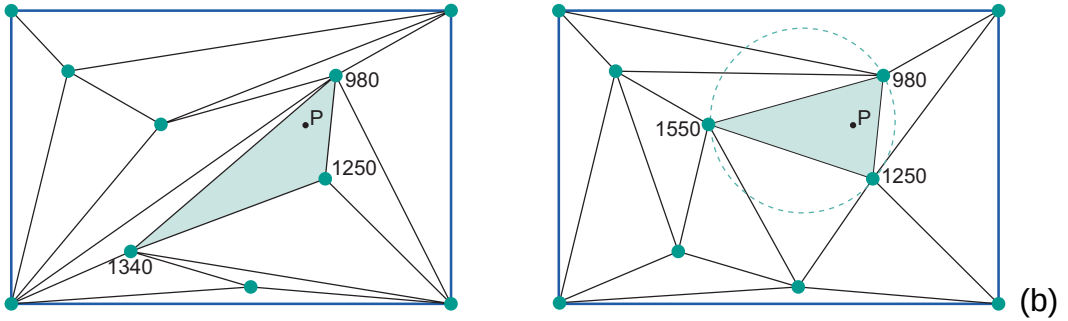


Figure 2.8: Two triangulations based on the input locations of Figure 2.7. (a) one with many 'stretched' triangles; (b) the triangles are more equilateral; this is a Delaunay triangulation.

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. So, it is wise to restrict the use of a plane to the triangular area 'between' the three points.

If we restrict the use of a plane to the area between its 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 2.8 demonstrates with two of them. Some tessellations are better than others, in the sense that they make smaller errors of elevation approximation. For instance, if we base our elevation computation for location P on the left hand shaded triangle, we will get another value than from the right hand shaded triangle. The second will provide a better approximation because the average distance from P to the three triangle anchors is smaller.

The triangulation of Figure 2.8(b) happens to be a *Delaunay triangulation*,

The gradient is a steepness measure indicating the maximum rate of elevation change, indicated as a percentage or angle. The aspect is an indication of which way the slope is facing; it can be defined as the compass direction of the gradient. More can be found in Section 4.5.3.

which in a sense is an optimal triangulation. There are multiple ways of defining what such a triangulation is [53], but we suffice here to state two important properties. The first is that the triangles are as equilateral ('equal-sided') as they can be, given the set of anchor points. The second property is 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.

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 tiling of the entire study space. The cells of this tiling, however, do not have an associated stored value as is typical of tessellations, but rather a simple interpolation function that uses the elevation values of the three anchor points.

Point representations

Points are defined as single coordinate pairs (x, y) when we work in 2D or coordinate triplets (x, y, z) when we work in 3D. The choice of coordinate system is another matter, and we will come back to it in Chapter 4.

Points are used to represent objects that are best described as shape- and sizeless, single-locality features. Whether this is the case really depends on the purposes of the spatial application and also on the spatial extent of the objects compared to the scale applied in the application. For a tourist city map, parks will not usually be considered as point features, but perhaps museums will be, and certainly public phone booths could be represented as point features.

Besides the georeference, usually extra data is stored for each point object. This so-called *administrative* or *thematic data*, can capture anything that is considered relevant about the object. For phone booth objects, this may include the owning telephone company, the phone number, the data last serviced *et cetera*.

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 application of mapping tourist information, bus, subway and streetcar 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.

At the beginning of Section 2.2, we saw that arbitrary, continuous curvilinear features are 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 define a *line*. Another word for internal node is *vertex* (plural: vertices); another phrase for line that is used in some GISs is polyline, arc or edge. A node or vertex is like a point (as discussed above) but it only serves to define the line; it has no special meaning to the application other than that.

The vertices of a line help to shape it, and 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 simple 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 only one.

Still, there are cases in which we would like to have the opportunity to use arbitrary curvilinear features as representation of real-world phenomena. Think of garden design with perfect circular or elliptical lawns, or of detailed topographic maps representing roundabouts and the annex sidewalks. All of this can be had in GIS in principle, but many systems do not at present accommo-

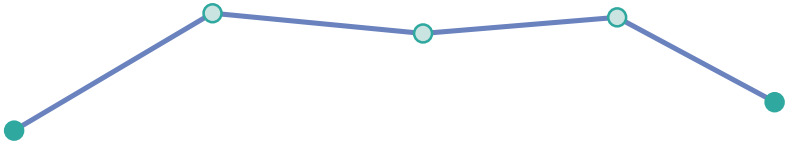


Figure 2.9: A line is defined by its two end nodes and zero or more internal nodes, also known as vertices. This line representation has three vertices, and therefore four line segments.

date such shapes. If a GIS supports some of these curvilinear features, it does so using parameterized mathematical descriptions. But a discussion of these more advanced techniques is beyond the purpose of this text book.

Collections of (connected) lines may represent phenomena that are best viewed as networks. With networks, specific type of interesting questions arise, that have to do with connectivity and network capacity. Such issues come up in traffic monitoring, watershed management and other application domains. With network elements—i.e., the lines that make up the network—extra values are commonly associated like distance, quality of the link, or carrying capacity.

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. In essence, what we then get is an application-determined (i.e., adaptive) partition of space, similar to, but not quite like an irregular tessellation of the raster approach.

Observe that a polygon representation for an area object is yet another example of a finite approximation of a phenomenon that inherently may have a curvilinear boundary. In the case that the object can be perceived as having a fuzzy boundary, a polygon is an even worse approximation, though potentially the only one possible.

An example is provided in Figure 2.10. It illustrates a simple study with three area objects, 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 simply 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. But this is far from optimal.

To understand why this is the case, take a closer look at the shared boundary between the bottom left and right polygons in Figure 2.10. The line that makes up the boundary between them is the same, which means that under the above representation it would be stored twice, namely once for each polygon. This is a form of data duplication—known as *data redundancy*—which turns out to be

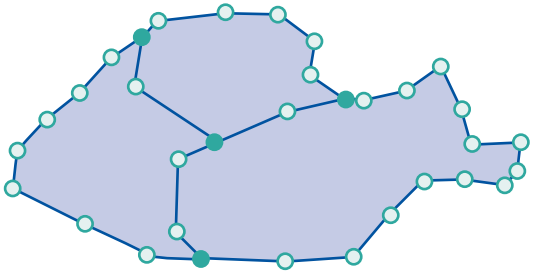


Figure 2.10: Areas as they are represented by their boundaries. Each boundary is a cyclic sequence of line features; each line—as before—is a sequence of two end nodes, with in between, zero or more vertices.

awkward in data maintenance.

There is another disadvantage to such polygon-by-polygon representations. If we want to find out which polygons border the bottom left polygon, we have to do a rather complicated and time-consuming analysis comparing the vertex lists of all boundary lines with that of the bottom left polygon. In the case of Figure 2.10, with just three polygons, this is fine, but when our data set has 5,000 polygons, with perhaps a total of 25,000 boundary lines, even the fastest computers will take their time in finding neighbour polygons.

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 is provided in Figure 2.11. It illustrates which additional information is stored about spatial relationships between lines and polygons, for instance. Obviously, real coordinates for nodes (and vertices) will also be stored, albeit in another table.

The boundary model is sometimes also called the topological data model as it captures some topological information, such as polygon neighbourhood. Observe that it is a simple query to find all the polygons that are the neighbour

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	...

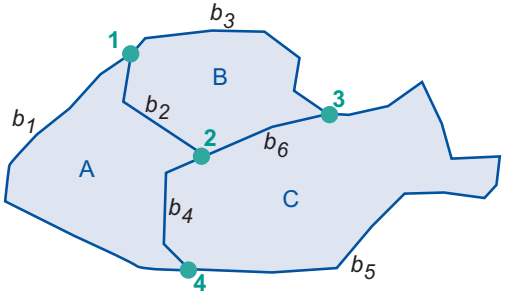


Figure 2.11: A simple boundary model for the polygons A, B and C. For each arc, we store the start and end node (as well as a vertex list, but these have been omitted from the table), its left and right polygon. The 'polygon' W denotes the outside world polygon.

of some given polygon, unlike the case we discussed above. We look at some of the topological issues in the next section.

2.2.4 Topology and spatial relationships

General spatial topology

Topology deals with spatial properties that do not change under certain transformations. A simple example will illustrate what we mean.

Assume you have some features that are drawn on a sheet of rubber (as in Figure 2.12). Now, take the sheet and pull on its edges, but do not tear or break it. The features will change in shape and size. Some properties, however, do 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, and
- the areas are still bounded by the same boundaries, only the shapes and lengths of their perimetry have changed.

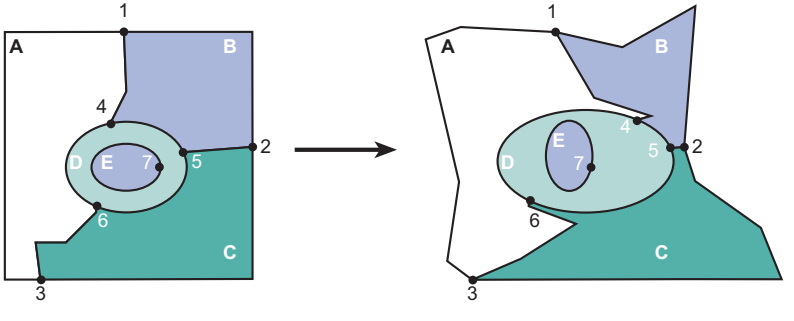


Figure 2.12: Rubber sheet transformation: The space is transformed, yet if we do not 'tear' or 'break', many relationships between the constituents remain unchanged.

These relationships are invariant under a continuous transformation. Such properties are called topological properties, and the transformation is called a *topological mapping*.

The mathematical properties of the geometric space used for spatial data can be described as follows.

- The space is a three-dimensional *Euclidean space* where for every point we can determine its three-dimensional coordinates as a triple (x, y, z) of real numbers. In this space, we can define features like points, lines, polygons, and volumes as geometric primitives of the respective dimension. A point is zero-dimensional, a line one-dimensional, a polygon two-dimensional, and a volume is 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*, of which the definition 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.

- *Interior* and *boundary* 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.

There are a number of advantages when our computer representations of geographic phenomena have built-in sensitivity of topological issues. Questions related to the 'neighbourhood' of an area are a point in case. To obtain some 'topological sensitivity' simple building blocks have been proposed with which more complicated representations can be constructed:

- We can define within the topological space features 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*. Figure 2.13 provides examples.

As the topological characteristics of simplices are well-known, we can infer the

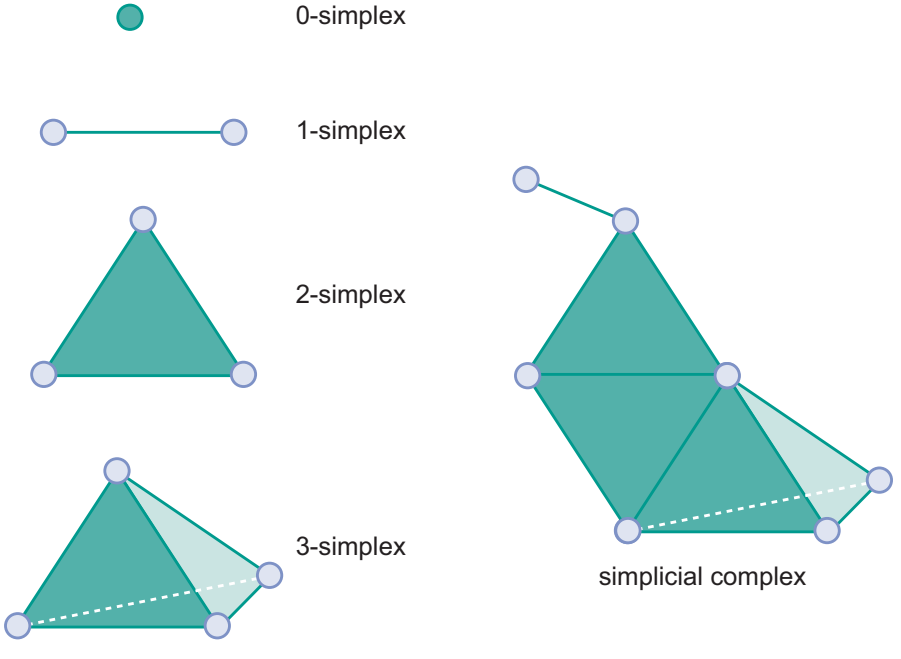


Figure 2.13: Simplices and a simplicial complex. Features are approximated by a set of points, line segments, triangles, and tetrahedrons.

topological characteristics of a simplicial complex from the way it was constructed.

The topology of two dimensions

We can use the topological properties of interior and boundary to define relationships between spatial features. Since the properties of interior and boundary do not change under topological mappings, we can investigate their possible relations between spatial features.[4] We can define the *interior* of a region R as the maximal set of points in R for which we can construct a disk-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 but that do not belong to the interior of R, i.e., one cannot construct a disk-like environment around such points that still belongs to R completely.

Suppose we consider a spatial region A. It has a boundary and an interior, both seen as (infinite) sets of points, and 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 a 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 as

$$\begin{aligned} A \; meets \; B \quad \stackrel{\text{def}}{=} \quad & interior(A) \cap interior(B) = \emptyset \; \wedge \\ & boundary(A) \cap boundary(B) \neq \emptyset \; \wedge \\ & interior(A) \cap boundary(B) = \emptyset \; \wedge \end{aligned}$$

[4] We restrict ourselves here to relationships between spatial *regions* (i.e., two-dimensional features without holes).

$$boundary(A) \cap interior(B) = \emptyset.$$

In the above formula, the symbol $\wedge$ expresses the logical connective 'and'. Thus, it states four properties that must all be met.

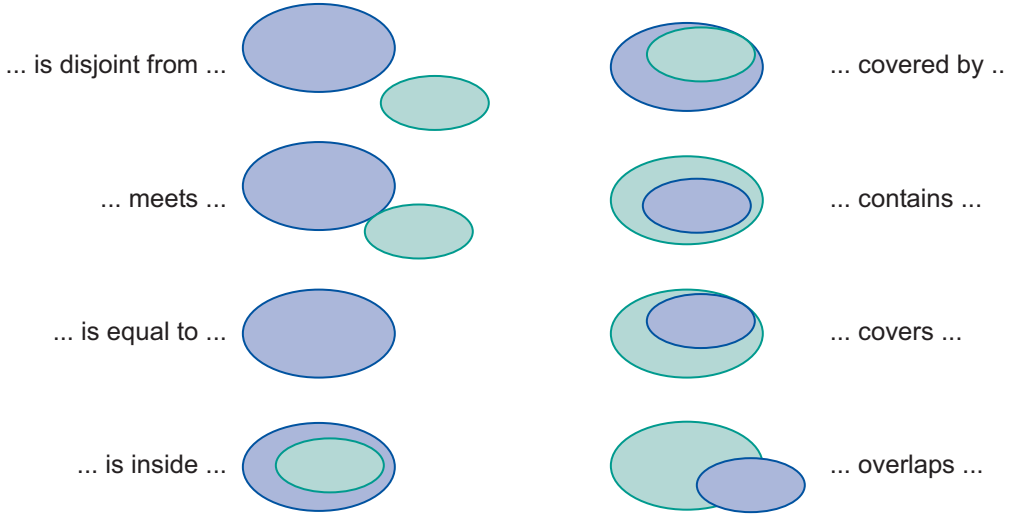


Figure 2.14: Spatial relationships between two regions derived from the topological invariants of intersections of boundary and interior. The relationships can be read with the green region on the left . . . and the blue region on the right . . .

Figure 2.14 shows all eight spatial relationships: *disjoint*, *meets*, *equals*, *inside*, *covered by*, *contains*, *covers*, and *overlaps*. These relationships can be used, for instance, in queries against a spatial database.

It turns out that the rules of how simplices and simplicial complexes can be emdedded in space are quite different for two-dimensional space than they are for three-dimensional space. Such a set of rules defines the *topological consistency* of that space. It can be proven that if the rules below are satisfied for all features in a *two*-dimensional space, the features define a topologically consistent configuration in 2D space. The rules are illustrated in Figure 2.15.

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 alternating (and cyclic) sequence of 1- and 2-simplices.
5. 1-simplices only intersect at their (bounding) nodes.

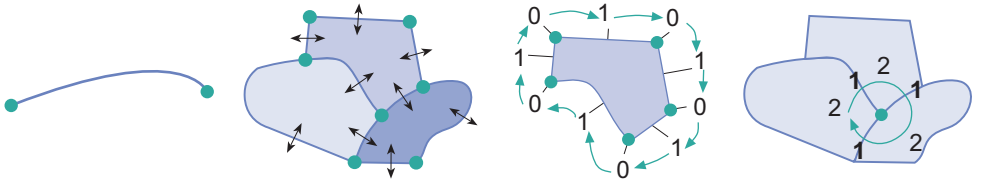


Figure 2.15: The five rules of topological consistency in two-dimensional space

The three-dimensional case

It is not without reason that our discussion of vector representations and spatial topology has focused mostly on objects in two-dimensional space. The history of spatial data handling is tainted almost purely 2D, and this is true also for the majority of present-day GIS applications. Still, quite a few application domains require elevational data as well, but these are usually accommodated by so-called $2\frac{1}{2}$D data structures.

These $2\frac{1}{2}$D data structures are similar to the (above discussed) 2D data structures using points, lines and areas. They also apply the rules of two-dimensional topology, as they were illustrated in Figure 2.15. This means that different lines cannot cross without intersecting nodes, and that different areas cannot overlap.

There is, on the other hand, one important aspect in which $2\frac{1}{2}$D data does 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, and therefore, a piecewise planar, 'wrinkled surface' can be constructed as well, much like a TIN. Note however, that one cannot have two different nodes with identical x- and y-coordinate, but different z-value. Consequently, true solids cannot be represented in a $2\frac{1}{2}$D GIS.

Solid representation is an important feature for some dedicated GIS application domains. Two of them are worth mentioning here: mineral exploration, where solids are used to represent ore bodies, and urban modelling, where solids may represent various human constructions like buildings and sewer canals. The three-dimensional characteristics of such objects are fundamental as 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

GIS is formed by the *polyhedra*, which are the solids limited by 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; this happens to be the case for the 3-simplex. Most polyhedra have many more facets; the cube has already six.

2.2.5 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 defy an easy answer sometimes.

Map scale can be defined as the ratio between distance on a paper map and 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 large, so typically it means there is much detail; 'small-scale' in contrast means a small ratio, hence fewer detail. When applied to spatial data, the term *resolution* is commonly associated with the cell width of the tessellation applied.

Digital spatial data, as stored in a GIS, is essentially without scale: scale is a ratio notion associated with visual output, like a map, not with the data that was used to produce the map. We will later see that digital spatial data can be obtained by digitizing a paper map (Section 4.1.2), and in this context we might informally say that the data is at this-and-that scale, indicating the scale of the map from which the data was derived.

2.2.6 Representations of geographic fields

In the above we have looked at various representation techniques. Now we can study which of them can be used to represent a geographic field.

A geographic field can be represented through a tessellation, through a TIN or through a vector representation. The choice between them is determined by the requirements of the application at hand. 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. We provide an example of the other two below.

Tessellation to represent a field

In Figure 2.16, we illustrate how a raster represents a continuous field like elevation. Different shades of blue indicate different elevation values, with darker blues indicating higher elevations. The choice of a blue colour spectrum is only to make the illustration aesthetically pleasing; real elevation values are stored in the raster, so instead we could have printed a real number value in each cell. This would not have made the figure very legible, however.

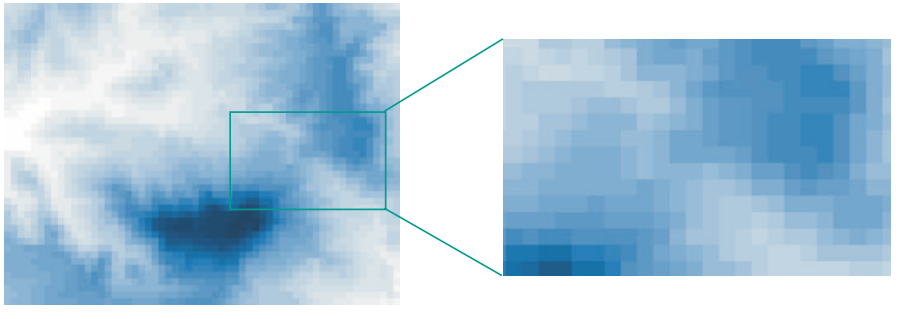


Figure 2.16: A raster representation (in part) of the elevation of the study area of Figure 2.2. Actual elevation values are indicated as shades of blue. The depicted area is the north-east flank of the mountain in the south-east of the study area. The right-hand side of the figure is a zoomed-in part of that of the left.

A raster can be thought of as a long list of field values: actually, there should be $m \times n$ such values. The list is preceded with some extra information, like a single georeference as the origin of the whole raster, a cell size indicator, the integer values for m and n, and a data type indicator that informs about how to interpret cell values. Rasters and quadtrees do not store the georeference of each cell, but infer it from the above information *about* the raster.

A TIN is a much 'sparser' data structure: the amount of data stored is less if we try to obtain a structure with approximately equal interpolation error, as compared to a regular raster. 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 are a certain guarantee for correct peaks and mountain slope faces.

Vector representation of a field

We shortly mention a final representation for fields like elevation, but with a vector flavour. This technique uses isolines of the field. An *isoline* is a linear feature that connects the points with equal field value. When the field is elevation, we also speak of *contour lines*. The elevation of the Falset study area is represented with contour lines in Figure 2.17. Both TINs and isoline representations use vectors.

Figure 2.17: A discretized elevation field representation for the study area of Figure 2.2. Indicated are elevation isolines at a resolution of 25 metres. Data source: Division of Engineering Geology (ITC)

Isolines as a *representation mechanism* are not very common, however. They are in use as a *geoinformation visualization technique* (in mapping, for instance), but commonly using a TIN for this type of field is the better choice. Many GIS packages allow to generate an isoline visualization from a TIN.

2.2.7 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 2.1.4), and many of these parameters can be expressed in terms of vectors.

Tessellations are not entirely out of the picture, though, and are commonly used for representing geographic objects as well.

Tessellations to represent geographic objects

Remotely sensed images are an important data source for GIS applications. Unprocessed digital images contain pixels, with each pixel 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 attempts to characterise each pixel into one of a finite list of classes, thereby obtaining an interpretation of the contents of the image. The classes recognized can be crop types as in the case of Figure 2.18 or urban land use classes as in the case of Figure 2.19. These figures illustrate the unprocessed images (a) as well as a classified version of the image (b).

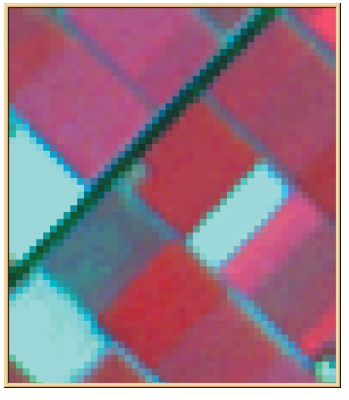
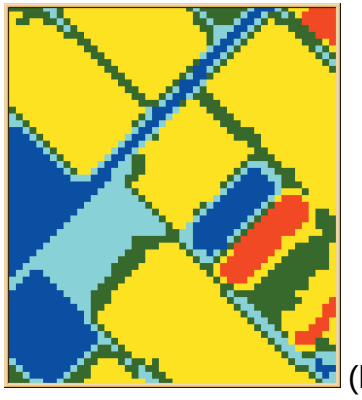
(a)   (b)

Figure 2.18: An unprocessed digital image (a) and a classified raster (b) of an agricultural area.

The application at hand may be interested only in geographic objects such as potato fields (Figure 2.18(b), in yellow) or industrial complexes (Figure 2.19(b), in red). This would mean that all other classes are considered unimportant, and are probably dropped from further analysis. If that further analysis can be carried out with raster data formats, then there is no need to consider vector

representations.

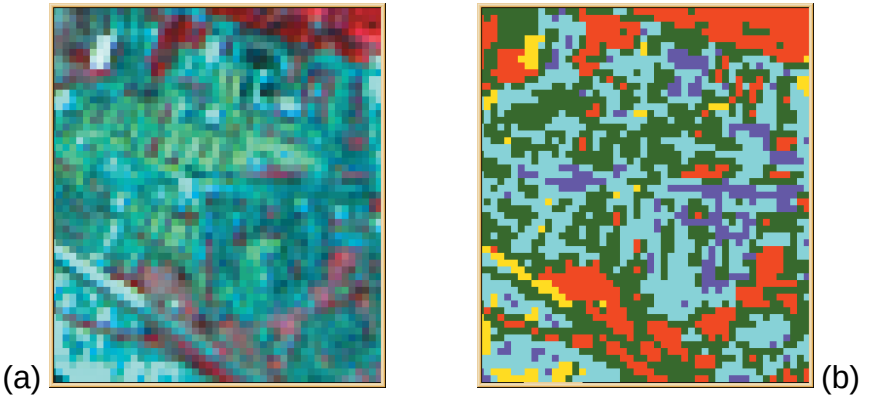


Figure 2.19: An unprocessed digital image (a) and a classified raster (b) of an urban area.

How the process of image classification takes place is not the subject of this book. It is dealt with extensively in *Principles of Remote Sensing* [30].

Nonetheless, we must make a few observations regarding the representation of geographic objects in rasters. *Area objects* are conveniently represented in raster, albeit that area boundaries may appear ragged. This is a typical by-product of raster resolution versus area size, and artificial cell boundaries. One must be aware, for instance, of the consequences for area size computations: what is the precision with which the raster defines the object's size?

Line and *point objects* are more awkward to represent using rasters. After all, we could say that rasters are area-based, and geographic objects that are perceived as lines or points are perceived to have zero area size. Standard classification techniques, moreover, may fail to recognise these objects as points or lines.

Many GIS do offer support for line representations in raster, and operations

on them. Lines can be represented as strings of neighbouring raster cells with equal value, as is illustrated in Figure 2.20. Supported operations are connectivity operations and distance computations. There is again an issue of precision of such computations.

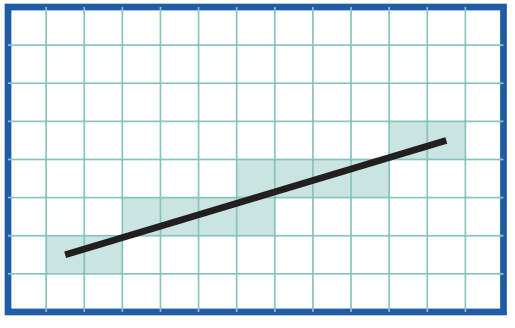


Figure 2.20: An actual straight line (in black) and its representation (light green cells) in a raster.

Vector representations for geographic objects

The somehow more natural way to represent geographic objects is by vector representations. We have discussed most issues already in Section 2.2.3, and a small example suffices at this stage.

Figure 2.21: Various objects (buildings, bike and road lanes, railroad tracks) represented as area objects in a vector representation.

In Figure 2.21, a number of geographic objects in the vicinity of the ITC building have been depicted. These objects are represented as area representations in a boundary model. Nodes and vertices of the polylines that make up the object's boundaries are not illustrated, though they obviously are stored.

2.3 Organizing one's spatial data

In the previous sections, we have discussed various types of geographic information and ways of representing them. We have looked at case-by-case examples, however, without looking much at how various sorts of spatial data are combined in a single system.

The main principle of *data organization* applied in GIS systems is that of a spatial data layer. A *spatial data layer* is either a representation of a continuous or discrete field, or a collection of objects of the same kind. The intuition is that the data is organized by kind: all telephone booth point objects would be in a single data layer, all road line objects in another one. A data layer contains spatial data—of any of the types discussed above—as well as attribute (or: thematic) data, which further describes the field or objects in the layer. Attribute data is quite often arranged in tabular form, as we shall see in Chapter 3. An example of two field data layers is provided in Figure 2.22.

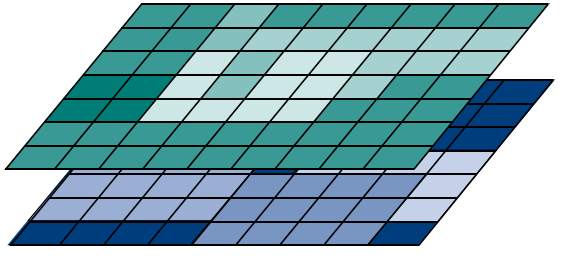


Figure 2.22: Different rasters can be overlaid to look for spatial correlations.

Data layers can be overlaid with each other, inside the GIS package, so as to study combinations of geographic phenomena. We shall see later that a GIS can be used to study the *spatial correlation* between different phenomena: in what way are occurrences/events occurring in the same location? To that end, a com-

putation is performed that overlays one data layer with another. This is schematically depicted in Figure 2.23 for two different object layers. But GIS software also allows to overlay field layers, or even a field with an object layer.

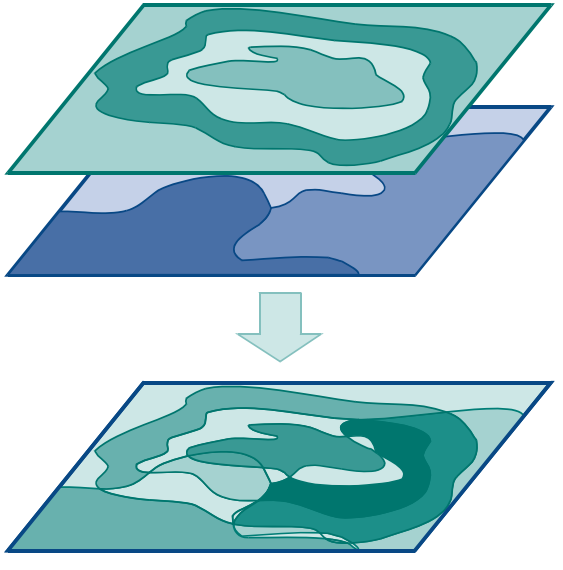


Figure 2.23: Two different object layers can be overlaid to look for spatial correlations, and the result can be used as a separate (object) layer.

In Chapter 3, we will look more into the functions offered by GISs, as well as by database systems.

2.4 The temporal dimension

2.4.1 Spatiotemporal data

Beside having geometric, thematic and topological properties, geographic phenomena change over time; we say that they have temporal characteristics. And for many applications, it is change over time that is quite often the most interesting aspect of the phenomenon to study. This area of work is commonly known as *change detection*. It is, for instance, interesting to know who were the owners of a land parcel in 1980, or how land cover changed from the original primary forest to pastures over time. Change detection addresses such questions as:

- Where and when did change take place?
- What kind of change occurred?
- With what speed did change occur?
- What else can be understood about the pattern of change?

The support that GISs offer for change detection is at present not very impressive. Most studies require substantial efforts from the GIS user in data preparation and data manipulation. Part of an example data set from such a project is provided in Figure 2.24. The purpose of this study was to assess whether radar images are reliable resources for detecting the disappearance of primary forests [7]. Typical for studies of this type, is the definition of a 'model of change', which includes knowledge and hypotheses of how change occurs. In this case, it included knowledge about speed of tree growth, for instance.

Spatiotemporal data structures are representations of geographic phenomena changing over time. Several representation techniques have been proposed in the literature. The most important ones will be discussed briefly below.

Observe that besides 2D or 3D space, the extra dimension of time is again inherently of a continuous nature, and that again, if we want to represent this in a computer, we will have to 'discretize' this dimension.

Before we describe the major characteristics of various techniques, we need a framework to describe the nature of time itself. The time dimension can be characterized with the following properties:

Time density Time can be measured along a *discrete* or *continuous* scale. Discrete time is composed of discrete elements (seconds, minutes, hours, days, months, or years). In continuous time, no such discrete elements exist, and for any two different points in time, there is always another point in between. We can also structure time by *events* (points in time) or *periods* (time intervals). When we represent time periods by a start and end event, we can derive temporal relationships between events and periods such as 'before', 'overlap', 'after', *et cetera*.

Dimensions of 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. Observe that the time at which we store something in the database/GIS typically is (much) later than when the related event took place.

Often, what we record in a computer system is a 'snapshot state' that represents a single point in time of an ongoing natural or man-made process. We may store a string of 'snapshot states' but must be aware that this is still only a feeble representation of that process.

Time order Time can be considered to be *linear*, extending from the past to the present ('now'), and into the future. 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 a week are recognized, make more sense and can be useful.

Measures of time When measuring time, we speak of a *chronon* as the shortest non-decomposable unit of time that is supported by a GIS or database (e.g., this could be a millisecond). The life span of an object is measured by a (finite) number of chronons. *Granularity* is the precision of a time value in a GIS or database (e.g., year, month, day, second, etc.). Different applications require different granularity. In cadastral applications, time granularity could well be a day, as the law requires deeds to be date-marked; in geological mapping applications, time granularity is more likely in the order of thousands or millions of years.

Time reference Time can be represented as *absolute* (*fixed time*) or *relative* (*implied time*). 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 may be relative to an arbitrary point in time.).

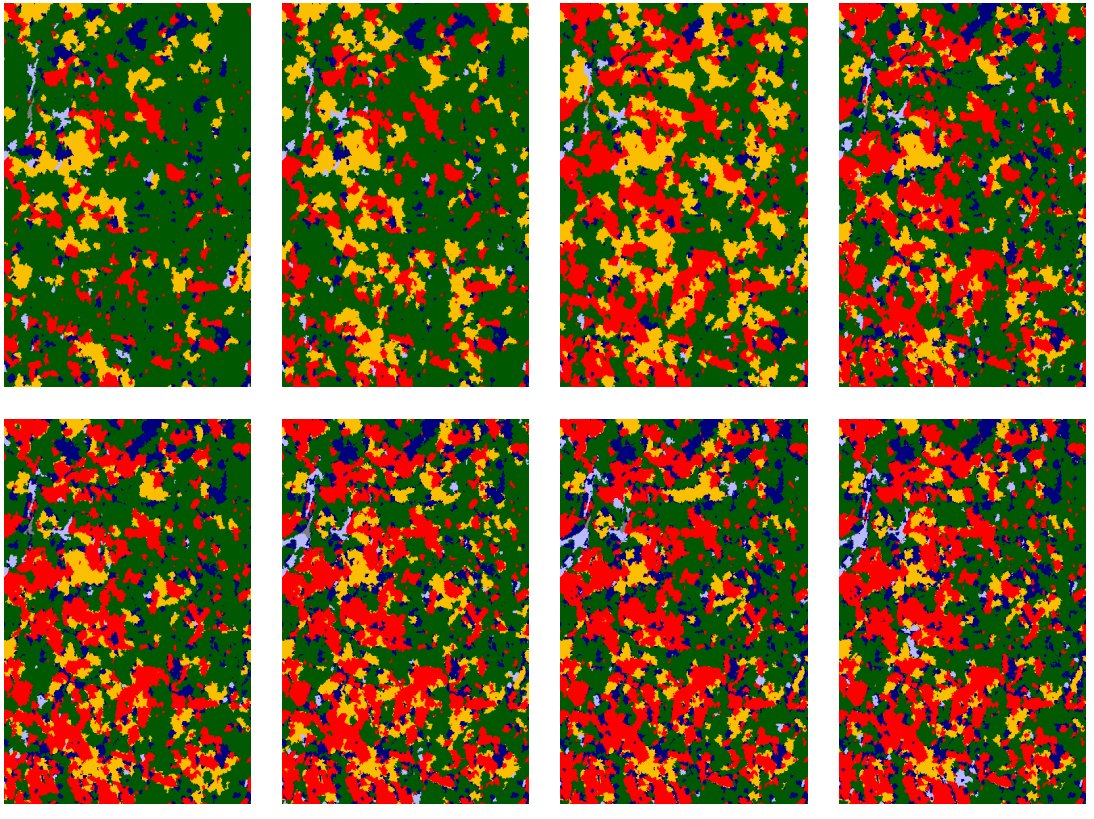


Figure 2.24: The change of land cover in a 9 × 14 km study site near San José del Guaviare, div. Guaviare, Colombia, during a study conducted in 1992–1994 by Bijker [7]. A time series of ERS–1 radar images after application of (1) image segmentation, (2) rule-based image classification, and (3) further classification using a land cover change model. The land cover classes are:
primary forest,
secondary vegetation,
secondary vegetation with *Cecropia* trees,
pasture, and
pasture & secondary vegetation.
Data source: Wietske Bijker, ITC.

2.4.2 Spatiotemporal data models

In spatiotemporal data models, we consider changes of spatial and thematic attributes over time. In data analysis, we can keep the *spatial domain fixed* and look only at the attribute changes over time for a given location in space. We would, for instance, be interested how land cover changed for a given location or how the land use changed for a given land parcel over time, provided its boundary did not change. Much of our discussion here and below is based on Langran's work [39].

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 could be interested to see which locations were covered by forest over a given period.

Finally, we can assume both the *spatial and attribute domain variable* and consider how fields or objects changed over time. This may lead to notions of *object motion*, and these are a subject of current research, with two of the applications being traffic control and mobile telephony. But many more applications are on the horizon: think of wildlife tracking, vector-borne disease control, and weather forecasting. Here, the problem of *object identity* becomes apparent. When does a change or movement cause an object to disappear and become a new one? With wildlife this is quite obvious; with weather systems less so. But this should no longer surprise as much: we have already seen that some geographic phenomena can perfectly well be described as objects, while others are better represented as fields.

In the following, we describe the main characteristics of some spatiotemporal data models.

The snapshot model

In the snapshot model, data layers for the same information theme are time-stamped. A data layer represents that state of affairs for the (valid) time with which it is time-stamped. This valid time is a specific extra attribute associated with the data layer. We do not have any information about the events that caused changes between the different states represented by layers. This model is based on a linear, absolute, and discrete time. It supports only valid time but can have variable time granularity. The spatial domain is fixed (and is typically field-based) and the attribute domain is variable.

As many current GISs lack support for temporal data, the snapshot model is the most commonly used model. GIS end-users, however, have to build their (time-stamped) data layers themselves, and commonly the GIS has no built-in awareness of time issues. This means that analysis of change manifested in the sequence of states is the complete responsibility of the end-user.

The snapshot model is the most common one in the Earth sciences, as satellite imagery is such an important base data source for them. After image classification of several images of the same area, we essentially have obtained a field-based snapshot sequence that might function as a basis for study with time-related questions.

The space-time cube model

Like the previous one, this model is based on a two-dimensional view of the study space (spanned by the x- and y-axis), in which geographic phenomena are traced through time (along the t-axis) thereby creating a three-dimensional space-time cube. A space-time cube represents a process in two-dimensional space, played out along a third, temporal dimension. The trace of some object through time creates a worm-like trajectory in the space-time cube. This model potentially allows absolute, continuous, linear, branching and cyclic time. It supports only valid time. The attribute domain is kept fixed and the spatial domain typically varies.

Given that current GISs already have a hard time ensuring data integrity even in the standard, atemporal case, it is somewhat difficult to forecast whether they will soon be capable of handling data integrity in space-time cube models. Topological correctness for a vector data layer can be achieved, but to ensure it under single object changes requires effort. Multiple, concurrent object changes are even more difficult to guard topologically, and the rules of full topological consistency under continuous change are not even well-understood.

The space-time cube model can be viewed as a idealized snapshot model with an infinitely dense snapshot sequence.

The space-time composite model

The space-time composite model also starts from a two-dimensional view of the study space at a given start time. Every change of an object that happens later is projected onto the initial data layer and is intersected with the existing features. This leads to successive intersections, thereby creating an incrementally built, finer polygon mesh. Over time, more and more polygons will be stored in the data layer. Every polygon in this mesh has its attribute history stored with it. The space-time composite model is based on linear, discrete, and relative time. It supports both valid and transaction time, and multiple granularity. It keeps the attribute domain fixed and the spatial domain variable.

This model can be useful if the amount of changes is limited, and changes are discrete steps, as is the case for instance in cadastral applications, where parcels may be split or joined. Even here, it may be wise to consider hybrid solutions. A commonly applied technique is to regularly 'start anew' with a new data layer with initially non-split polygons.

The event-based model

In an event-based model, we start with an initial state and record events along the time line. Whenever a change occurs, an entry is recorded. This is a time-based model. The spatial and thematic attribute domains are secondary. The model is based on discrete, linear, relative time, and supports only valid time and multiple granularity.

Our event records on the event-based model are such that we can reconstruct the full spatial and non-spatial history of our study area. This reconstruction will require some or much computation. This, therefore, is a model with low storage consumption but with high costs in computation.

Summary

In this chapter, we have taken a closer look at different types of geographic phenomena, and looked into the ways of how these can be represented in a computer system, such as a GIS. Geographic phenomena are present in the real world that we study; their computer representations only live inside computer systems.

We found that an important distinction between phenomena is whether it is omnipresent—i.e., occurring everywhere in the study area—or whether its constituents somehow 'sparsely' populate the study area. The first class of phenomena we called fields, the second class objects.

Amongst fields, we identified continuous and discrete phenomena. Continuous phenomena could even be differentiable, meaning that for locations factors such as gradient and aspect can be determined.

Amongst objects, important classification parameters turned out to be location, shape, size and orientation. We further saw that a fundamental part of the shape parameter is the dimension of the object: is it a point, line, area or volume object? In all cases, it turns out that a representation of the boundary of the object, whether crisp or fuzzy, is often used in GIS.

In the second half of the chapter, we elaborated on the techniques with which the above phenomena are actually stored in a computer system. The fundamental problem in obtaining realistic representations is that these are usually continuous in nature, thus requiring an infinite data collection to represent them faithfully. As a consequence of the finite memory that we have available in computer systems, we must accept finite representations. This leads to approximations, and therefore error, in our GIS data.

Questions

1. For your own GIS application domain, make up a list of at least 20 different geographic phenomena that might be relevant in the application domain.

2. On page 69, we defined geographic phenomena as representable as triplets, and discussed three phenomena for the El Niño example of Chapter 1. For each of these, provide a to-the-point description of the triplets involved.

3. Take the list of question 1 and identify which phenomena are fields and which are objects. Which of your example objects are crisp?

4. There exists an obvious natural relationship between remotely sensed images and geographic fields, as we have defined them in this chapter, yet the two are not the same thing. Elaborate on this, and discuss what are the differences.

5. Location, shape, size and orientation are potentially relevant characteristics of geographic objects. Try to provide an application example in which these characteristics do make sense for (a) point objects, (b) line objects, (c) area objects.

6. On page 71, we stated a rule-of-thumb, namely that natural phenomena are more often fields, whereas man-made phenomena are more often objects. Provide counter-examples from your favourite GIS application domain to this rule: name at least one natural phenomenon that is better perceived as object(s), and name a man-made phenomenon that is better perceived as field. (The latter is more difficult.)

7. On page 106, we provided the (logical) definition of the 'meets' relationship. Provide your version of the definitions of 'covered by' and 'overlaps'. Explain why this set of topological relationships between regions is also known as the *four-intersection scheme*.

8. What colour is the northwest quadrant of the outermost northeast quadrant of Figure 2.6? First check the field on the left, then use the quadtree on the right. What colour is the southeast quadrant of the outermost northeast quadrant?

9. Compute the area covered by green in the field of Figure 2.6, first by looking at the field, then by analysing the quadtree. What formula did you use?

10. Make an educated guess at the elevation of location P in Figure 2.7. What are the gradient and the aspect of the slope in this location, approximately? In a second stage, do this again, now based on the tessellations of Figure 2.8 (first the left one, then the right one).

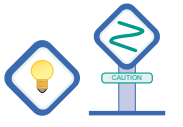
11. Explain how many line objects and how many line segments are illustrated in Figure 2.11. Complete the table on the left, using a numbering of vertices that you have made up yourself for Figure 2.10.

12. In the chapter, we have discussed raster-based and vector-based representations of geographic phenomena. We have not explicitly discussed what are the advantages and disadvantages of either. What do you think they are?

13. In Figure 2.20, we presented an actual line, and its representation in the raster. Compute the real length of the line (taking cell width as the unit). In rasters, when a GIS computes a distance it uses 1 as the distance between two cells that share a side, and it uses $\sqrt{2}$ as the distance between two cells that share only a corner point. What would be the computed length by the GIS of the line's representation in Figure 2.20? What can be said in general about the two lengths?

14. We have emphasized throughout the chapter that GIS representations of geographic phenomena are necessarily finite, notwithstanding the naturally continuous or curvilinear nature of the objects that we study. We are thus approximating, and are making errors by doing so. Do you think there is any way of *computing* what the errors are that we are making?

15. What observations can be made from a visual interpretation of Figure 2.24? What changes do you 'detect'? Which stages of change?



16. Discuss what type of time properties people are using in making appointments day by day.

Chapter 3

Data processing systems

Data processing systems are computer systems with appropriate hardware components for the processing, storage and transfer of data, as well as software components for the management of the hardware, peripheral devices and data. This chapter discusses the components of data processing systems that allow handling spatial data and derive geoinformation.

First, we discuss in brief some trends about computer hardware and software that have become apparent in recent years. These trends allow us to look ahead into the future and to attempt a forecast of what geoinformation processing may look like in ten years from now.[1]

Geographic information systems (GISs) as a tool for spatial data handling are

[1] Both terms *geoinformation processing* and *spatial data handling* are commonly used in the field of GIS, and mean more or less the same. The first emphasizes more the aspect of interpretation and human understanding of the data afterwards, whereas the latter emphasizes more the technical issues of how computers operate on the data that represent our geographic phenomena. We will use both terms liberally.

discussed next. We look at their general functions, but will not deal with them in detail, as these functions are highlighted extensively in Chapter 4 and 5. In Section 3.3, we discuss database management systems (DBMSs), including some principles of data extraction from a database, as that is not covered elsewhere in this book. We finalize with a section on the combined use of GIS and DBMS, namely Section 3.3.6.

3.1 Hardware and software trends

The developments in computer hardware proceed at an enormously fast speed. Almost every six months, a faster, more powerful processor generation replaces the previous one, and makes our computers an estimated 30% faster.

Computers get smaller and at the same time, their performance increases. The power that we have available in today's portable notebook computers is a multiple of the performance that the first PC had when it was introduced in the early 1980s. In fact, current PC systems have orders of magnitude more memory and storage than the so-called minicomputers of 20 years ago. Moreover, they fit on an office desk. At the same time, software providers produce application programs and operating systems that consume more and more memory. To efficiently run a computer with Windows 2000 and some general purpose office applications, a PC should be minimally equipped with 64 Mbytes of main memory and 4 or more Gbytes of disk storage, as we write this.

Software technology develops somewhat slower and often cannot fully use the possibilities offered by the hardware, but existing software obviously performs better when run on faster computers.

Also, computers have become increasingly portable. Hand-held computers are now commonplace in business and personal use. For a long time, the Achilles heel in computer portability—actually: in appliance portability—has been the weight and capacity of carry-on batteries. Breakthroughs are on their way for these as well. Portable computers will soon become common and cheap, allowing field surveyors, for instance, to take with them powerful computers into the field, possibly hooked up with GPS receivers for instantaneous georeferencing.

Another major development of recent years is in computer networks. In

essence, we have now arrived in an era where any computer can almost anywhere on Earth be hooked up onto some network, and contact other computers virtually anywhere else. This allows fast and reliable exchange of (spatial) data as well as of the computer programs to operate on them.

Mobile phones are frequently used to communicate with computers and the Internet. The communication between portable computers and networks is still rather slow when they are connected via a mobile phone. The transmission rate currently supported by mobile communication providers is only 9,600 bits per second (bps). Digital telephone links (ISDN) supports up to 64,000 bps, and high-speed computer networks have a capacity of several million bps. The new ADSL technology that is coming to the market now supports a rate of about 6 Mbps. With the upcoming arrival of UMTS (Universal Mobile Telecommunications System), digital communication of text, audio, and video becomes possible at a rate of approximately 2 Mbps. The combination of GPS receiver, portable computer and mobile phone is then one that may dramatically change our world, and certainly so for Earth science professionals with out-of-office activities.

Open systems use agreed upon, standard, architectures and protocols for networking. This makes it easier to link different systems. Interoperability is the ability of hardware and software of computers from different vendors to communicate with each other. An interoperable database would for instance allow differently formatted databases to appear as a single homogenous database to a user.

3.2 Geographic information systems

The handling of spatial data usually involves processes of data acquisition, storage and maintenance, analysis and output. For many years, this has been done using analogue data sources, manual processing and the production of paper maps. The introduction of modern technologies has led to an increased use of computers and digital information in all aspects of spatial data handling. The software technology used in this domain is geographic information systems.

Typical planning projects require data sources, both spatial and non-spatial, from different institutes, like mapping agency, geological survey, soil survey, forest survey, or the census bureau. These data sources may have different time stamps, and the spatial data may be in different scales and projections. With the help of a GIS, the maps can be stored in digital form in a database in world coordinates (metres or feet). This makes scale transformations unnecessary, and the conversion between map projections can be done easily with the software. The spatial analysis functions of the GIS are then applied to perform the planning tasks. This can speed up the process and allows for easy modifications to the analysis approach.

3.2.1 The context of GIS usage

Spatial data handling involves many disciplines. We can distinguish disciplines that develop spatial concepts, provide means for capturing and processing of spatial data, provide a formal and theoretical foundation, are application-oriented, and support spatial data handling in legal and management aspects. Table 3.1 shows a classification of some of these disciplines. They are grouped according to how they deal with spatial information. The list is not meant to be exhaustive.

The discipline that deals with all aspects of spatial data handling is called geoinformatics. It is defined as:

> Geoinformatics is the integration of different disciplines dealing with spatial information.

Geoinformatics has also been described as "the science and technology dealing with the structure and character of spatial information, its capture, its classification and qualification, its storage, processing, portrayal and dissemination, including the infrastructure necessary to secure optimal use of this information" [23]. Ehlers and Amer [19] define it as "the art, science or technology dealing with the acquisition, storage, processing production, presentation and dissemination of geoinformation."

A related term that is sometimes used synonymously with geoinformatics is geomatics. It was originally introduced in Canada, and became very popular in French speaking countries. Laurini and Thompson [40] describe it as "the fusion of ideas from geosciences and informatics." The term geomatics, however, was never fully accepted in the United States where the term geographical information science is preferred. Goodchild [22] defines GIS research as "research on the generic issues that surround the use of GIS technology, impede its successful

implementation, or emerge from an understanding of its potential capabilities."

Characteristics of disciplines	*Sample disciplines*
Development of spatial concepts	Geography Cognitive Science Linguistics Psychology
Means for capturing and processing spatial data	Remote Sensing Surveying Engineering Cartography Photogrammetry
Formal and theoretical foundation	Computer Science Expert Systems Mathematics Statistics
Applications	Archaeology Architecture Forestry Earth Sciences Regional and Urban Planning Surveying
Support	Legal Sciences Economy

Table 3.1: Disciplines involved in spatial data handling

3.2.2 GIS software

The main characteristics of a GIS software package are its analytical functions that provide means for deriving new geoinformation from existing spatial and attribute data. A GIS can be defined as follows [4]:

> A GIS is a computer-based system that provides the following four sets of capabilities to handle georeferenced data:
>
> 1. input,
> 2. data management (data storage and retrieval),
> 3. manipulation and analysis, and
> 4. output.

Depending on the interest of a particular application, a GIS can be considered to be a data store (i.e., a database that stores spatial data), a toolbox, a technology, an information source or a field of science (as part of spatial information science).

Like in any other discipline, the use of tools for problem solving is one thing, to produce these tools is something different. Not all tools are equally well-suited for a particular application. Tools can be improved and perfected to better serve a particular need or application. The discipline that provides the background for the production of the tools in spatial data handling is *spatial information theory*.

All GIS packages available on the market have their strengths and weaknesses, resulting typically from the package's development history and/or intended application domain(s). Some GIS have traditionally focused more on

support for raster manipulation, others more on (vector-based) spatial objects. We can safely state that any package that provides support for only raster or only objects, is *not* a full-fledged, generic GIS. Well-known, full-fledged GIS packages in use at ITC are ILWIS and ArcInfo. Both are in use in practical sessions of the core curriculum on GIS principles, which is why this text book tries to describe the field of GIS independent from them: the book must be useful to users of either package!

One cannot say that one GIS package is 'better' than another one: it all depends what one wants to use the package for. ILWIS's traditional strengths have been in raster processing and scientific spatial data analysis, especially suitable in what we called project-based GIS applications on page 44. ArcInfo has been renowned more for its support of vector-based spatial data and their operations, user interface and map production, a bit more typical of institutional GIS applications. Any such brief characterization, however, does not do justice to these packages, and it is only after extended use that preferences become clear.

3.2.3 Software architecture and functionality of a GIS

A geographic information system in the wider sense consists of software, data, people, and an organization in which it functions. In the narrow sense, we consider a GIS as a software system for which we discuss its architecture and functional components.

According to the definition, a GIS always consists of modules for input, storage, analysis, display and output of spatial data. Figure 3.1 shows a diagram of these modules with arrows indicating the data flow in the system. For a particular GIS, each of these modules may provide many or only few functions. However, if one of these functions would be completely missing, the system should not be called a geographic information system.

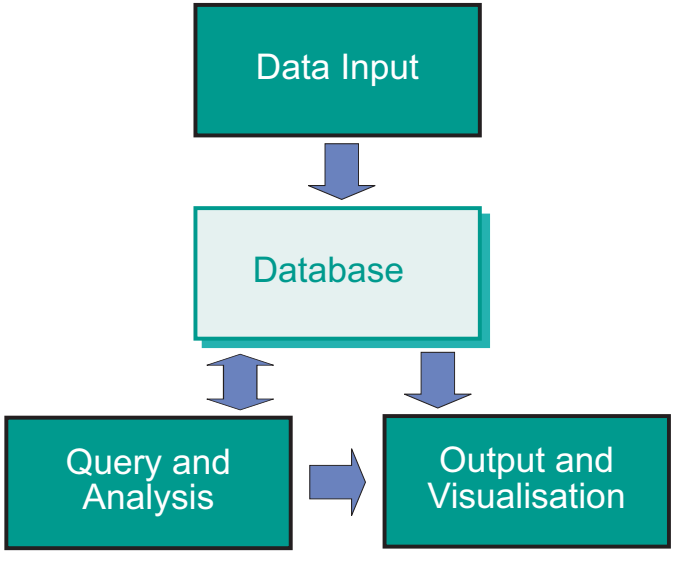


Figure 3.1: Functional components of a GIS

An explanation of the various functions of the four components for data input, storage, analysis, and output can provide a functional description of a GIS. Here, we only briefly describe them. A more detailed treatment can be found in follow-up chapters.

Beside data input (data capture), storage and maintenance, analysis and output, geoinformation processes involve also dissemination, transfer and exchange as well as organizational issues. The latter define the context and rules according to which geoinformation is acquired and processed.

Method	*Devices*
Manual digitizing	• coordinate entry via keyboard • digitizing tablet with cursor • mouse cursor on the computer monitor (heads-up digitizing) • (digital) photogrammetry
Automatic digitizing	• scanner
Semi-automatic digitizing	• line following devices
Input of available digital data	• magnetic tape or CD-ROM • via computer network

Table 3.2: Spatial data input methods and devices used

Data input

The functions for data input are closely related to the disciplines of surveying engineering, photogrammetry, remote sensing, and the processes of digitizing, i.e., the conversion of analogue data into digital representations. Remote sensing, in particular, is the field that provides photographs and images as the raw base data from which to obtain spatial data sets. Additional techniques for obtaining spatial data are *manual digitizing*, *scanning* and sometimes *semi-automatic line following*.

Today, digital data on various media and on computer networks are used increasingly. Table 3.2 lists the methods and devices used in the data input process. More discussion on spatial data input can be found in Chapter 4.

Method	*Devices*
Hard copy	• printer • plotter (pen plotter, ink-jet printer, thermal transfer printer, electrostatic plotter) • film writer
Soft copy	• computer screen (CRT)
Output of digital data sets	• magnetic tape • CD-ROM • computer networks

Table 3.3: Data output and visualization

Data output and visualization

Data output is closely related to the disciplines of cartography, printing and publishing. Table 3.3 lists different methods and devices used for the output of spatial data.

Cartography and scientific visualization make use of these methods and devices to produce their products. The importance of digital products (data sets) is increasing and data dissemination on digital media or on computer networks becomes extremely important. Chapter 6 is devoted to visualization techniques.

In both data input and data output, the Internet has a major share. The World Wide Web plays the role of an easy to use interface to repositories of large data sets. Aspects of data dissemination, security, copyright, and pricing require special attention. The design and maintenance of a *spatial information infrastructure* deals with these issues.

Data storage

The representation of spatial data is crucial for any further processing and understanding of that data. In most of the available processing systems, data are organized in layers according to different themes or scales. They are stored either according to thematic categories, like land use, topography and administrative subdivisions, or according to map scales, representing map series of different scale. An important underlying need or principle is a representation of the real world that has to be designed to reflect phenomena and their relationships as close as possible to what exists in reality.

In a spatial database, features are represented with their (geometric and non-geometric) attributes and relationships. The geometry of features is represented with (geometric) primitives of the respective dimension. These primitives follow either the vector or the raster approach.

As described in Chapter 2, 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) pieces, mostly a square tessellation of dimension two or three (these pieces are called *pixels* in 2D, *voxels* in 3D), and indicates for every piece which object it covers, in case it represents a discrete field. In case of a continuous field, the pixel holds a representative value for that field. Table 3.4 lists advantages and disadvantages of raster and vector representations.

Storing a raster, in principle, is a straightforward issue. A raster is stored in a file as a long list of values, one for each cell, preceded by a small list of extra information (the so-called file 'header') that informs how to interpret the list. The order of the cell values in the list can be—but need not be—left-to-right, top-to-bottom. This simple space filling scheme is known as *row ordering*, see Figure 3.2 (a). The header of the raster file will typically inform how many rows

Tessellation representation	*Vector representation*
advantages	*advantages*
• simple data structure • simple implementation of overlays • efficient for image processing	• efficient representation of topology
disadvantages	*disadvantages*
• less compact data structure • difficult to represent topology	• complex data structure • overlay more difficult to implement • inefficient for image processing

Table 3.4: Tessellation and vector representations compared

and columns the raster has, which space filling scheme is used, and what sort of values are stored for each cell.

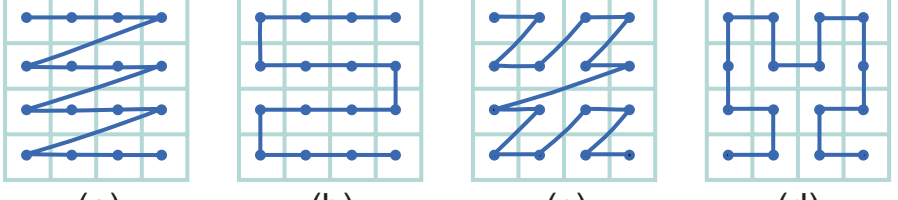


Figure 3.2: Four types of space filling curves: (a) row order, (b) row-prime order, (c) Morton (Z) order, (d) Peano-Hilbert order.

Other space filling schemes are illustrated in Figure 3.2 (b) to (d), in which the dark blue line indicates the order of cell values in the list. These schemes may seem to be overly complicated, but they have nice characteristics. The most important one of these is that compared to the row ordering scheme, the oth-

ers keep values of neighbouring cells closer together in the value list. This is important when one wants to extracting only a part of the raster from storage.

Low-level storage structures for vector data are much more complicated, and a discussion is certainly beyond the purpose of this introductory text. The best intuitive understanding can be obtained from Figure 2.11, where a boundary model for polygon objects was illustrated. Similar structures are in use for line objects. A fundamental consideration for the design of storage structures for any type of vector-based object is *spatial proximity*. In essence, it states that objects that are near in geographic space should be near in storage space as well. Fetching data from storage is done in units of a disk page, the smallest consecutive piece of stored data. The essence of spatial proximity will ensure that if we fetch one object from storage it is likely that its nearest neighbour objects are in the same disk page. For further, advanced reading we can suggest [57].

Spatial (vector) and attribute data are quite often stored in separate structures. Some sort of boundary model, as discussed above, is used for the spatial data, while the attribute data is stored in some tabular format. Typically, the vector objects in the first are given identifying values that the tables in the second use as reference. This is the way to link attribute with vector data. More detail on these issues is provided in Section 3.3.6.

GIS software packages provide support for both spatial and attribute data, i.e., they support spatial data storage using a vector approach, as well as attribute data support with tables. Historically, however, database management systems (DBMS) have been based on the notion of tables for data storage. Compared with what DBMS offer, GIS table functionality usually is not impressive. It is no surprise therefore that more and more GIS applications make use of a DBMS for attribute data support, while keeping the spatial data inside the GIS package. Most GISs nowadays allow to link with a DBMS and to exchange

attribute data with it. We will take a closer look at DBMS techniques in Section 3.3.1. But first, we focus on GIS functionality.

3.2.4 Querying, maintenance and spatial analysis

The most distinguishing part 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 satisfying user needs. The combination of a database, GIS software, rules, and a reasoning mechanism (implemented as a so-called inference engine) leads to what is sometimes called a *spatial decision support system* (SDSS).

In a GIS, data are stored in layers (or themes). Usually, several themes are part of a project. The analysis functions of a GIS use the spatial and non-spatial attributes of the data in a spatial database to answer questions about the real world.

In spatial analysis, various kinds of question may arise. They are listed with their possible answers and the required GIS functions in Table 3.5.

The following three classes are the most important query and analysis functions of a GIS, after [4]:

- Maintenance and analysis of spatial data,
- Maintenance and analysis of attribute data, and
- Integrated analysis of spatial and attribute data.

The first and third are GIS-specific, so are dealt with here. the second class is discussed in Section 3.3.

Questions	*Answers*	*GIS functions*
What is . . . ?	Display of data as maps, reports and tables, e.g., “What are the name and the address of the owner of that land parcel?”	Storage and query functions
What pattern . . . ?	Patterns in the data, e.g., all parcels with an area size greater than 2000.	Query functions with constraints
What . . . if . . . ?	A prediction about the data at a certain time or at a certain location.	Modelling functions

Table 3.5: Types of queries

Maintenance and analysis of spatial data

Maintenance of (spatial) data can best be defined as the combined activities to keep the data set up-to-date and as supportive as possible to the user community. It deals with obtaining new data, and entering them into the system, possibly replacing outdated data. The purpose is have available an up-to-date, stored data set. After a major earthquake, for instance, we may have to update our digital elevation model to reflect the current elevations better so as to improve our hazard analysis.

Operators of this kind operate on the spatial properties of GIS data, and provide a user with functions as described below.

Format transformation functions convert between data formats of different systems or representations, e.g., reading a DXF file into a GIS.

Geometric transformations help to obtain data from an original hard copy source through digitizing the correct world geometry. These operators transform device coordinates (coordinates from digitizing tablets or screen coordinates) into world coordinates (geographic coordinates, metres, etc.).

Map projections provide means to map geographic coordinates onto a flat surface (for map production), and *vice versa*.

Edge matching is the process of joining two or more map sheets. At the map sheet edges, feature representations have to be matched so as to be combined.

Graphic element editing allows to change digitized features so as to correct errors, and to prepare a clean data set for topology building.

Coordinate thinning is a process that often is applied to remove redundant vertices from line representations.

Integrated analysis of spatial and attribute data

Analysis of (spatial) data can be defined as computing from the existing, stored data set new information that provides insights we possibly did not have before. It really depends on the application requirements, and the examples are manifold. Road construction in mountainous areas is a complex engineering task with many cost factors such as the amount of tunnels and bridges to be constructed, the total length of the tarmac, and the volume of rock and soil to be moved. GIS can help to compute such costs on the basis of an up-to-date digital elevation model and soil map.

Functions of this kind operate on both spatial and non-spatial attributes of data, and can be grouped into the following types.

Retrieval, classification, and measurement functions

- Retrieval functions allow the selective search and manipulation of data without the need to create new entities.
- Classification allows assigning features to a class on the basis of attribute values or attribute ranges (definition of data patterns).
- Generalization is a function that joins different classes of objects with common characteristics to a higher level (generalized) class.[2]

[2]The term *generalization* has different meanings in different contexts. In geography the term 'aggregation' is often used to indicate the process that we call generalization. In cartography, generalization means either the process of producing a graphic representation of smaller scale from a larger scale original (*cartographic generalization*), or the process of deriving a coarser resolution representation from a more detailed representation within a database *(model generalization*). Finally, in computer science generalization is one of the *abstraction mechanisms* in object-orientation.

- Measurement functions allow measuring distances, lengths, or areas.

Overlay functions belong to the most frequently used functions in a GIS application. They allow to combine two spatial data layers by applying the set-theoretic operations of intersection, union, difference, and complement using sets of positions (geometric attribute values) as their arguments. Thus we can find

- the potato fields on clay soils (intersection),
- the fields where potato or maize is the crop (union),
- the potato fields not on clay soils (difference),
- the fields that do not have potato as crop (complement).

Neighbourhood functions operate on the neighbouring features of a given feature or set of features.

- Search functions allow the retrieval of features that fall within a given search window (which may be a rectangle, circle, or polygon).
- Line-in-polygon and point-in-polygon functions determine whether a given linear or point feature is located within a given polygon, or they report the polygons that a given point or line are contained in.
- The best known example of proximity functions is the buffer zone generation (or buffering). This function determines a fixed-width (or variable-width) environment surrounding a given feature.
- Topographic functions compute the slope or aspect from a given digital representation of the terrain (digital terrain model or DTM).

- Interpolation functions predict unknown values using the known values at nearby locations.
- Contour generation functions calculate contours as a set of lines that connect points with the same attribute value. Examples are points with the same elevation (contours), same depth (bathymetric contours), same barometric pressure (isobars), or same temperature (isothermal lines).

Connectivity functions accumulate values as they traverse over a feature or over a set of features.

- Contiguity measures evaluate characteristics of spatial units that are contiguous (are connected with unbroken adjacency. Think of the search for a contiguous area of forest of certain size and shape.
- Network analysis is used to compute the shortest path (in terms of distance or travel time) between two points in a network (routing). Alternatively, it finds all points that can be reached within a given distance or duration from a centre (allocation).
- Visibility functions are used to compute the points that are visible from a given location (viewshed modelling or viewshed mapping) using a digital terrain model.

3.3 Database management systems

A large, computerized collection of structured data is what we call a *database*. In the non-spatial domain, databases have been in use since the 1960s, for various purposes like bank account administration, stock monitoring, salary administration, order bookkeeping, and flight reservation systems. These applications have in common that the amount of data is usually quite large, but that the data itself has a simple and regular structure.

Setting up a database is not an easy task. One has to consider carefully what the database purpose is, and who will be its users. Then, 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*. After its design, we may start to enter data into the database. Of equal importance is keeping the data up-to-date, and it is usually wise to make someone responsible for regular maintenance of the database. Throughout the whole process it is essential to document all the design decisions made. Such documentation is crucial for an extended database life. Many enterprise databases tend to outlive the professional careers of their designers.

A *database management system* (DBMS) is a software package that allows the user to set up, use and maintain a database. Like A GIS allows to set up a GIS application, a DBMS offers generic functionality for database organization and data handling. Below, we will take a closer look at what type of functions are really offered by DBMSs. Many standard PCs are equipped these days with a DBMS called Access. This package is quite functional but only for smaller (private) databases.

In the next paragraphs, we will take a look at strengths and weaknesses of database systems (Section 3.3.1), and a standard for data structuring, called the

relational data model (Section 3.3.3). In between, Section 3.3.2 looks at our options when we decide *not* to use a DBMS for our data management, and discusses alternatives. Then, we discuss a technique for data extraction from a database (Section 3.3.4) and various aspects of recent database developments in Section 3.3.5.

3.3.1 Using a DBMS

There are various reasons why one would want to use a DBMS to support 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 use in practice. The result may be that finding simple facts takes minutes, and performing simple calculations perhaps even hours.

- A DBMS can be instructed to guard over some levels of *data correctness*.

For instance, an important aspect of data correctness is data entry checking: making sure that the data that is entered into the database is sensible data that does not contain obvious errors. Since we know in what study area we work, we know the range of possible geographic coordinates, so we can make the DBMS check them.

The above 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, and their definition is part of the development of a database.

- A DBMS supports the *concurrent use* of the same data set by many users.

Moreover, for different users of the database, different views of the data can be defined. In this way, users will be under the impression that they operate on their personal database, and not on one shared by many people. This DBMS function is called *concurrency control*.

Large data sets are built up over time, which means that substantial investments are required to create them, and that probably many people are involved in the data collection, maintenance and processing. These data sets are often considered to be of a high strategic value for the owner(s), which is why many may want to make use of them within an organization.

- A DBMS provides a high-level, *declarative query language*.[3]

The most important use of the language is the definition of queries. A *query* is a computer program that extracts data from the database that meet the conditions indicated in the query. We provide a few examples below.

- A DBMS supports the use of a *data model*. A data model 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*. We discuss it in full in Section 3.3.3. Its primitives are *tuples* (also known as records, or rows) with attribute values, and *relations*, being sets of similarly formed tuples.

- A DBMS includes *data backup* and *recovery* functions to ensure data availability at al times.

[3]The word 'declarative' means that the query language allows the user to define *what* data must be extracted from the database, but not *how* that should be done. It is the DBMS itself that will figure out how to extract the data that is requested in the query. Declarative languages are generally considered user-friendlier because the user need not care about the 'how' and can focus on the 'what'.

As potentially many users rely on the availability of the data, the data must be safeguarded against possible calamities. Regular back-ups of the data set, and automatic recovery schemes provide an insurance against loss of data.

- A DBMS allows to control *data redundancy*.

A well-designed database takes care of storing single facts only once. Storing a fact multiple times—a phenomenon known as *data redundancy*—easily leads to situations in which stored facts start to contradict each other, causing reduced usefulness of the data. Redundancy, however, is not necessarily always an evil, as long as we tell the DBMS where it occurs so that it can be controlled.

3.3.2 Alternatives for data management

A good question at this point is whether there are any alternatives to using a DBMS, when one has a data set to care about. Obviously, it all depends on how much data there is or will be, what type of use we want to make of it, and how many people will be involved.

On the small-scale side of the spectrum—when the data set is small, its use relatively simple, and with just one user—we might use simple text files, and a text processor. Think of a personal address book as an example, or a not-too-big batch of simple field observations.

If our data set is still small and numeric by nature, and we have a single type of use in mind, perhaps a spreadsheet program will do the job. This can be the case if we have a number of field observations with measurements that we want to prepare for statistical analysis. However, if 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. Spreadsheets also do not accommodate multiple uses of the same data set well.

All too often, we find that data collections—if they are made digital—reside in text files or spreadsheets, when the type(s) of use that the owner has in mind really requires a DBMS. Text files offer no support for data analysis whatsoever, except perhaps alphabetical ordering. Spreadsheets do support some data analysis, especially when it comes to calculations over a single table, like averages, sums, minimum and maximum values. All of such computations are, however, restricted to just a single table of data. When one wants to relate the values in the table with values of another nature in some other table, an expert hand and an effort in time are usually needed. It is precisely here where the knowledge of a good database query language pays off.

3.3.3 The relational data model

A data model is a language that allows the definition of

- the *structures* that will be used to store the base data,
- the *integrity constraints* that the stored data has to obey at all moments in time, and
- the *computer programs* used to manipulate the data.

For the relational data model, the structures are attributes, tuples and relations to define the database structure. The computer programs either perform data extraction from the database without altering it, in which case we call them *queries*, or they change the database contents, and we speak of *updates* or *transactions*.

Let us look at a tiny database example from a cadastral setting. It is illustrated in Figure 3.3. This database consists of three tables, one for storing private people details, one for storing parcel details and a third one 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 parcels and the date of a title deed (DeedDate). The technical terms surrounding database technology are introduced below.

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 3.3: A small example database consisting of three relations (tables), all with three attributes, and resp. three, four and four tuples. 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.'

Relations, tuples and attributes

In the relational data model, a database is viewed as a collection of *relations*, commonly also known as *tables*. 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, as in Figure 3.3, relations can be displayed as tabular form data.

An *attribute* is a named field of a tuple, with which each tuple associates a value, the tuple's *attribute value*. All tuples in the same relation must have the same named attributes. They need, obviously, not have the same value for these attributes. The example relations provided in the figure should clarify this. The PrivatePerson table has three tuples; the Surname attribute value for the first tuple illustrated is 'Garcia.'

The phrase 'similarly shaped tuples' is taken a little bit further. It requires that the tuples do not only have the same attributes, but also 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 the set of integer number values, the set of real number values, et cetera. In our example cadastral database, the domain of the Surname attribute, for instance, is 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, and

PrivatePerson	(TaxId : string, Surname : string, Birthdate : date)
Parcel	(Pid : number, Location : polygon, AreaSize : number)
TitleDeed	(Plot : number, Owner : string, DeedDate : date)

Table 3.6: The relation schemas for the three tables of the database in Figure 3.3.

3. what the domain of each attribute is.

A relation definition obtained in this way is known as the *relation schema* of that relation. The definition of relation schemas is an important part of database design. Our example database has three relation schemas; one of them is TitleDeed. The relation schemas together make up the database schema. For the database of Figure 3.3, the relation schemas are given in Table 3.6. 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 only rarely change over time. This is not true of the tuples stored in tables: they, typically, are often changing, either because new tuples are added, others are removed, or yet others will see 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 3.3 gives us a single *database instance*, i.e., one relation instance for each relation. One relation instance 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

A well-designed database stores accessible information. The stored tuples represent facts of interest. What is interesting or relevant—and thus, what are the stored facts—depends on the purpose of the database. In our cadastral database, the facts concern the ownership of parcels. Typical factual units are parcels, title deeds and private people. Hence, we identified the three distinct relations.

Remember that we stated that database systems are particularly good at storing large quantities of data. One may think of perhaps tens of thousands of tuples per table. (Our example database is not even small, it is tiny!) To find any tuple in a really large table is almost impossible through a visual check. The DBMS must support quick searches amongst many tuples. This is why the relational data model uses the notion of key.

A *key* of a relation comprises one or more attributes. A value for these attributes uniquely identifies a tuple. In other words, if we have a value for each of the key attributes we are guaranteed to find at most 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, however, is 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. Thus, we want a key to have the fewest possible number of attributes.

If a key has just one attribute, it obviously can not have fewer 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

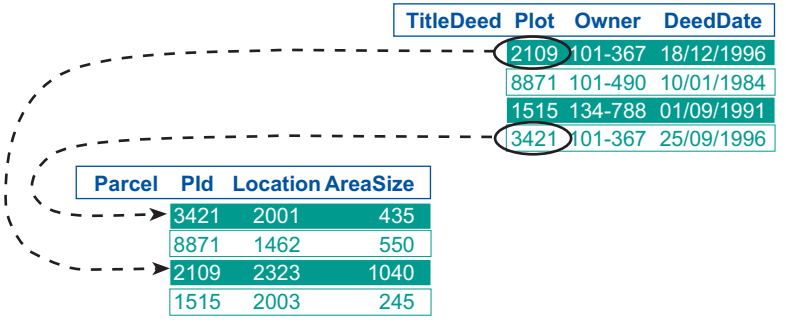


Figure 3.4: The table TitleDeed has a foreign key in its attribute Plot. This attribute refers to key values of the Parcel relation, as indicated for two TitleDeed tuples. The table TitleDeed actually has a second foreign key in the attribute Owner, which refers to PrivatePerson tuples.

single plot (in case of plots that are sold often) but also many title deeds for a single person (in case of wealthy persons).

As an aside, remark that an attribute such as AreaSize in relation Parcel is *not* a key, although it appears to be one in Figure 3.3. The reason is that some day there could be a second parcel with size 435, giving us two parcels with that value.

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 3.4.

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, therefore, is not a key of the relation in which it appears, despite its name!

Observe that a foreign key must have as many attributes as the primary key that it refers to.

The three golden rules of data integrity

A DBMS can be set up to guard over the correctness of the data that it stores. Data correctness is also known as data integrity. Intimately connected with the relational data model are three golden rules of data integrity that any database instance must adhere to. We have already seen the first rule, and it is called *Key uniqueness*.

Key uniqueness the key value of any tuple in any relation instance must be different from that of any other tuple in the same relation instance.

This rule speaks for itself: keys are meant to be unique identifiers, so duplicate primary key values are not allowed.

Key integrity the value of any key attribute of any tuple in any relation instance is always known.

We are not allowed to leave such values 'blank'.[4]

Observe that we stated "in any relation instance." This rule, like the first, should never be violated: not in yesterday's database, our current database or tomorrow's.

Referential integrity the value of a foreign key is either 'blank' (for all its attributes), or it is the key value of an existing tuple in the relation that the foreign key refers to.

One can think of referential integrity along the lines of a telephone directory, which provides the telephone numbers of people. If, for some person, no number is provided (represented as a 'blank' value in a database), we assume that

[4]The correct term here is 'null value', but a full discussion is beyond the purpose of this text.

person has no telephone. If, however, a number is provided, we assume that that number is correct. In other words, the telephone directory should give no number or a correct number.

3.3.4 Querying a relational database

We will now look at the three most elementary data extraction operators. They are quite powerful because they can be combined to define queries of higher complexity.

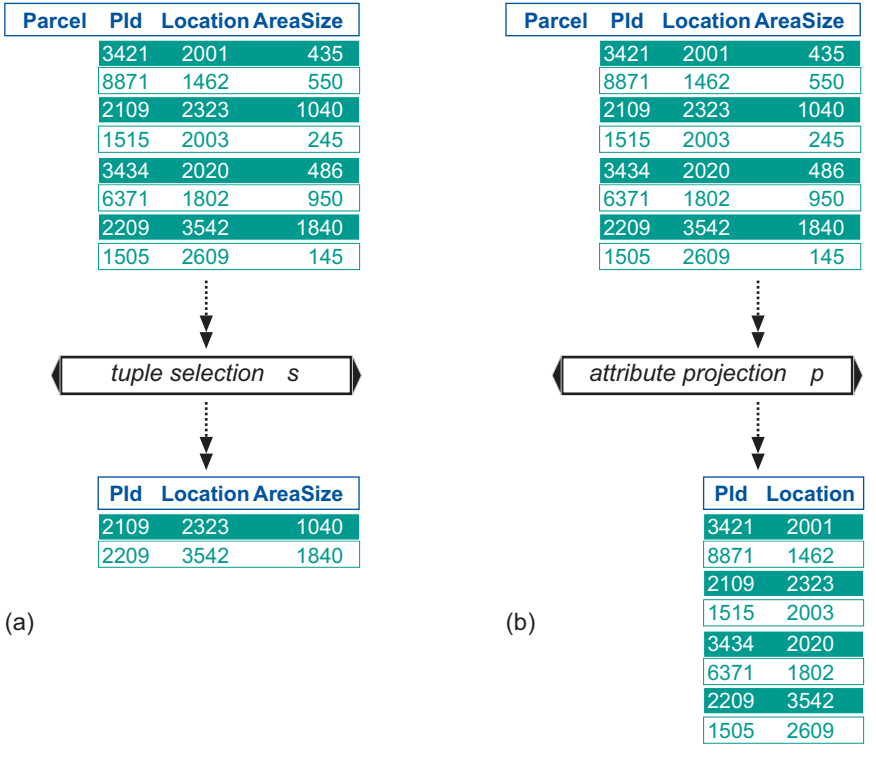


Figure 3.5: The two unary query operators: (a) tuple selection has a single table as input and produces another table with less tuples. Here, the condition was that AreaSize must be over 1000; (b) attribute projection has a single table as input and produces another table with fewer attributes. Here, the projection is onto the attributes PId and Location.

The three query operators have some features in common. 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, and this gives us the possibility to build more and more complex queries, if we want.

The first query operator is called *tuple selection*; it is illustrated in Figure 3.5(a), and works as follows. 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, 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.

An important observation is that the tuple selection operator produces an output relation with the same schema as the input relation, but with fewer tuples.

A second operator is also illustrated in Figure 3.5. It is called *attribute projection*. 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, and 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 way of defining queries in a relational database is through the *SQL* language. SQL stands for Structured Query Language. The two queries of Figure 3.5 are written in SQL as follows:

```
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.

Queries like the two above do not automatically 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 stored, however, for future use. The user can re-execute the query again to obtain a view on the result once more.

Our third query operator differs from the two above as it requires two input relations instead of one. The operator is called the *join*, and is illustrated in Figure 3.6. The output relation of this operator has as attributes those of the first and those of the second input relation. The number of attributes therefore increases. The output tuples are obtained by taking a tuple from the first input relation and 'gluing' it with 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 3.6 combines TitleDeed tuples with Parcel tuples, but only those for which the foreign key Plot matches with primary key PId.

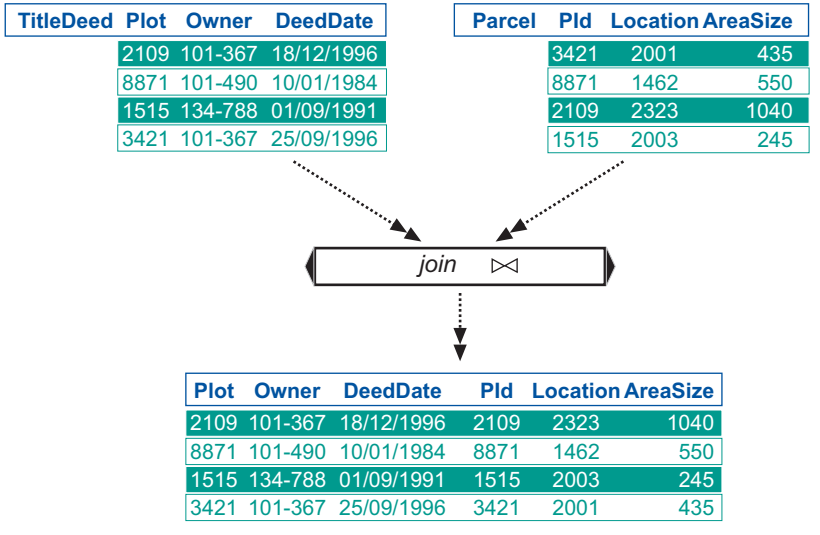


Figure 3.6: The essential binary query operator: join. The join condition for this example is TitleDeed.Plot=Parcel.Pid, which expresses a foreign key/key link between TitleDeed and Parcel. The result relation has $3 + 3 = 6$ attributes.

The above join query is also easily expressed in SQL as follows.

```
SELECT  *
FROM    TitleDeed, Parcel
WHERE   TitleDeed.Plot = Parcel.PId
```

The FROM-clause identifies the two input relations; the WHERE-clause states the join condition.

It is often not sufficient to use just one query for extracting sensible information from a database. The strength of these operators hides in the fact that they can be combined to produce interesting query definitions. We provide a final example to illustrate this. Take another look at the join of Figure 3.6. 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 above join, 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 3.7.

Finally, we may look at the SQL statement that would give us the query of Figure 3.7. It can be written as

```
SELECT   Owner, DeedDate
FROM     TitleDeed, Parcel
WHERE    TitleDeed.Plot = Parcel.PId AND AreaSize > 1000
```

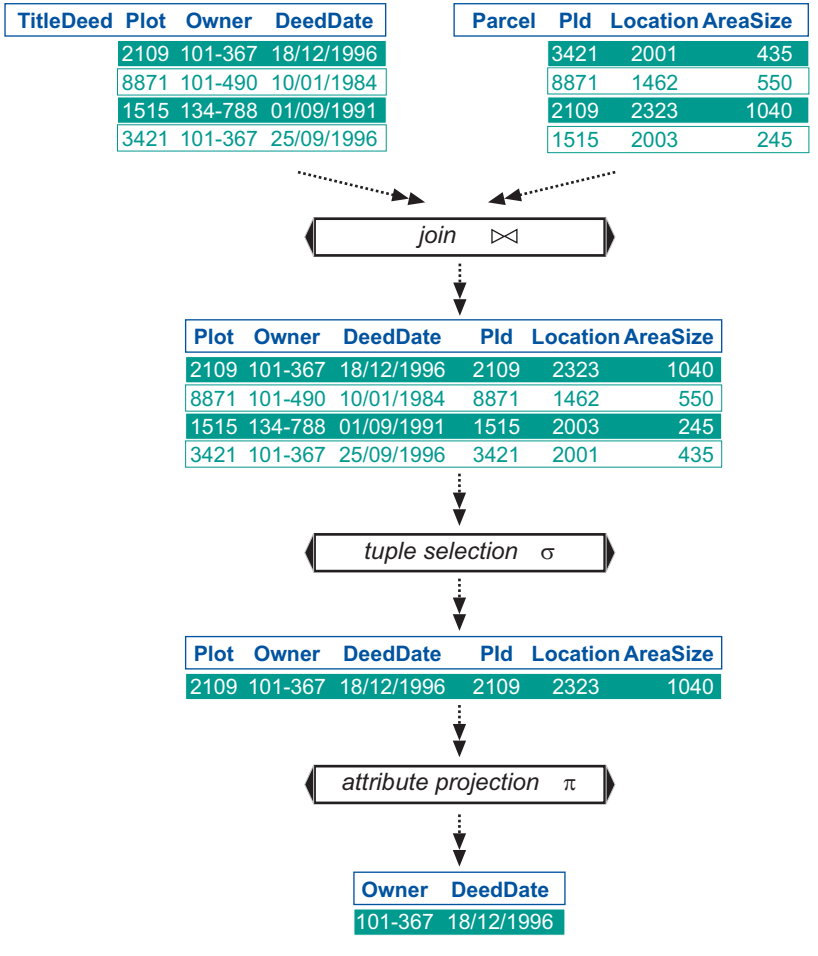


Figure 3.7: A combined selection/projection/join query, selecting owners and deed dates for parcels with a size larger than 1000. The join is carried out first, then follows a tuple selection on the result tuples of the join. Finally, an attribute projection is carried out.

3.3.5 Other DBMSs

The relational databases for which we provided examples above were first built in the early 1970s. They are a commercial success story because their use allowed many institutes and companies to build and maintain large administrative systems to support their information management. Relational databases are particularly good for standard administrative purposes like stock control, personnel administration, account management et cetera. All of these applications can be characterized as voluminous in terms of the amount of data, yet simple in terms of the type of data.

Relational databases are not very good at storing more complex types of data. In particular, and from a geographic perspective, they are not set up well to deal with spatial data. This is not to say they are useless for this purpose, but there is definitely room for improvement.

DBMS vendors have over the last 15 years recognized that need also and have developed data models beyond the relational data model. The most important general data models in this category are *object-oriented* and *object-relational* data models. We mention them here for completeness sake, and refer the interested reader to introductions as in [16, 20].

DBMS vendors have also understood the needs from various application fields, which has resulted in the development of various add-on packages to their DBMSs. One can now buy extensions for time series data management, internet support, spatial data, multimedia, financial data *et cetera*. It is to be expected that large, data-intensive GIS applications will soon start relying fully on the DBMS support for spatial data.

3.3.6 Using GIS and DBMS together

GIS and DBMS packages have developed in different directions, addressing different purposes. Yet, both store data and allow the user to manipulate the data to produce, hopefully relevant, results.

DBMSs have a long tradition in handling attribute (i.e., administrative, non-spatial, tabular, thematic—we use these terms interchangeably) data in a secure way, for multiple users at the same time. Some of the data in GIS applications is attribute data, so it makes sense using a DBMS for it. GIS packages themselves can store tabular data as well, however, they do not always provide a full-fledged query language to operate on the tables.

The strength of GIS technology lies in the built-in 'understanding' of geographic space and all functions that derive from it: spatial data structures for storage, spatial data analysis, and map production, for instance. Most GIS do not accommodate multi-user access naturally. We have also discussed above that DBMSs now start offering support for *spatial* data storage. Clearly, many choices must be made in setting up a GIS application.

The future is probably that large-scale GIS applications will require the use of both: DBMS for data storage (and multi-user support), GIS for spatial functionality. In such a setting, the DBMS will serve as a centralized data repository for all users, while each user would run her/his own GIS that obtains its data from the DBMS. Small-scale GIS applications, on the other hand, may not require a DBMS, and can be supported by a stand-alone GIS package.

In the section below, we look at current practice and situations in which GIS and DBMS are combined.

Attribute data in GIS applications

A GIS uses the raster and vector approach for representing geographic phenomena, but it must also record descriptive information about these phenomena. It does this typically in an attribute database subsystem. This in turn requires that the GIS must provide a link between the spatial data represented with rasters or vectors, and their non-spatial attribute data. These links turn the GIS into a special system: the user can store and examine information about *where* things are and *what* they are like, and such investigations can be bi-directional, from spatial data to attribute data and *vice versa*.

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 3.8 indicates the land use class for each of its cells, while an accompanying table provides full descriptions for all classes, including perhaps some statistical information for each of the types. Observe the analogy with the key/foreign key concept in relational databases.

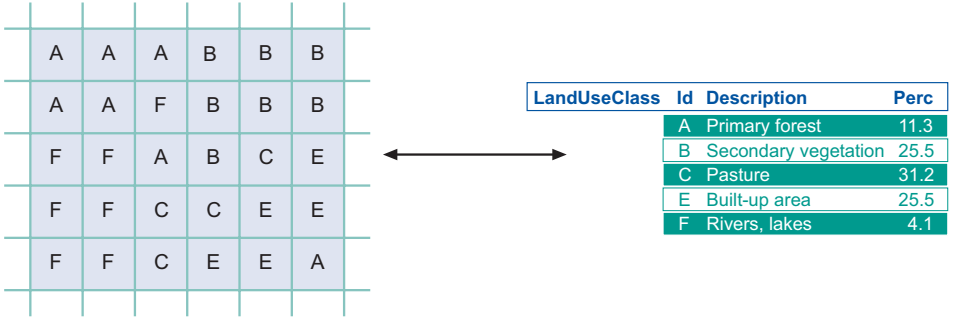


Figure 3.8: A raster representing land use and a related table providing full text descriptions (amongst others) of each land use class.

With vector representations, our spatial objects—whether they are points, lines or polygons—will be given a unique identifier by the system automatically.

This identifier is usually just called the object's 'ID' and can be used to link the spatial object (as represented in vectors) with its attribute data in an attribute table. The principle applied here is similar to that in raster settings, but now 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. Obviously, several tables may make such references to the vector system, but it is not uncommon to have some main table for which the ID is actually also the key.

There is, however, not always such an obvious one-to-one correspondence between the spatial data and the attribute data. For instance, consider the case where a long-term hydrological field survey includes daily rainfall measurements for many stations. It is to be expected that we would have one spatial data layer that represents the stations as point objects. In addition, one or more tables will be used to store the daily measurements, which over time will build up in volume. With any single station, we will have many measurements associated, and thus, the relationship between attribute data (the measurements) and spatial data (the stations) is many-to-one.

Depending on the computational requirements of our hydrological analysis model, we may have to perform various selections, joins and arithmetic or statistical computations with the measurement data, before we want to relate back to the station(s). It is only after these computations that we relate the attribute data with the spatial data.

The database tables mentioned above could have been stored within the GIS or in a separate DBMS. Smaller projects may do the first, but larger projects or those with higher computational requirements typically do the second. Present-day GIS packages allow to initialize the system such that the data exchange with an external DBMS is not too difficult. The details of this vary amongst packages.

Summary

In this chapter, we have made a tour of two brands of software systems that help in organizing our spatial and attribute data. We have seen that a GIS is more suited for the first and a DBMS is better for the second purpose. Yet, in spatial applications we usually have both kinds of data, so we must know both types of technology.

Many GISs allow to store and manipulate attribute data. They do so in two different ways. The oldest is to provide a little on-board database subsystem that offers some DBMS functionality but not all that one would expect. This is fine for applications of a more isolated or smaller character, but it is dangerous if the system to be built will have to support a larger user audience. This can be the case in bigger organizations or longer-term projects. Then, the second way becomes the more natural to follow. It involves using a full-fledged DBMS next to the GIS, and letting the DBMS handle all the attribute data. Many GISs nowadays provide software interfaces to external DBMS systems, so that the two can communicate their data.

We have tried to provide an overview and typology of the possibilities that GIS and DBMS technology in combination have to offer. But a full understanding of these possibilities will only be achieved after hands-on experience.

Questions

1. Consider the hypothetical case that your institute or company equips you for field surveys with a GPS receiver, a mobile phone (global coverage) and portable computer. Compare that situation with one where your employer only gives you a notepad and pencil for field surveying. What is the gain in time efficiency? What sort of project can be contemplated now that was impossible before?

2. Table 3.2 lists various ways of getting digital data into a GIS. From a perspective of data accuracy and data correctness, what do you think are the best choices? In your field, what is the commonest technique currently in use? Do you feel better techniques may be available?

3. In your domain of geoinformation application, provide examples of each of the query types listed in Table 3.5.

4. Although this chapter does not specifically describe what is meant by the terms, try to define what entails 'edge matching' and 'coordinate thinning' as mentioned on page 160. If possible, make a drawing that explains the principles. Consider what must be done to the spatial data.

5. Take a closer look at Figure 3.2 on page 155. Choose one of the four central cells in the raster as object of study, and determine the *average distance* along the space filling curve from the chosen cell to its eight neighbour cells. Do so for all four curves. What do you find? How is the situation for a cell in the middle of the left edge?

6. In Figure 3.3 and Table 3.6 we illustrated the structure of our example database. In what (fundamental) way does the table differ from the figure? Why have the attributes been grouped the way they have? (Hint: look for the obvious explanation.)

7. The following is a correct SQL query on the database of Figure 3.3. Explain in words what information it will produce when executed against that database.

```
SELECT PrivatePerson.Surname, TitleDeed.Plot
FROM   PrivatePerson, TitleDeed
WHERE  PrivatePerson.TaxId = TitleDeed.Owner AND
       PrivatePerson.BirthDate > 1/1/1960
```

   Determine what table the query will result in. If possible, draw up a diagram like Figure 3.6 (but without showing data values) that demonstrates what the query does.

Chapter 4

Data entry and preparation

The first step of using a GIS is to provide it with data. The acquisition and preprocessing of spatial data is an expensive and time-consuming process. Much of the success of a GIS project, however, depends on the quality of the data that is entered into the system, and thus this phase of a GIS project is critical and must be taken seriously.

Spatial data can be obtained from various sources. We discuss a number of these sources in Section 4.1. The specificity of spatial data obviously lies in it being spatially referenced. An introduction to spatial reference systems and related topics is therefore provided in Section 4.2. Issues concerning data checking and clean-up, multi-scale data, and merging adjacent data sets are discussed in Section 4.3. Section 4.4 provides an overview of preparation steps for point data. Several methods used for point data interpolation are elaborated upon. The use of elevation data and the preparation of a digital terrain model is the topic of the optional Section 4.5.

4.1 Spatial data input

Spatial data can be obtained from scratch, using direct spatial data acquisition techniques, or indirectly, by making use of spatial data collected earlier, possibly by others. Under the first heading fall field survey data and remotely sensed images. Under the second fall paper maps and available digital data sets.

4.1.1 Direct spatial data acquisition

The primary, and sometimes ideal, way to obtain spatial data is by direct observation of the relevant geographic phenomena. This can be done through ground-based field surveys *in situ*, or by using remote sensors in satellites or airplanes. An important aspect of ground-based surveying is that some of the data can be interpreted immediately by the surveyor. Many Earth sciences have developed their own survey techniques, and where these are relevant for the student, they will be taught in subsequent modules, as ground-based techniques remain the most important source for reliable data in many cases.

For remotely sensed imagery, obtained from satellites or aerial reconnaissance, this is not the case. These data are usually not fit for immediate use, as various sources of error and distortion may have been present at the time of sensing, and the imagery must first be freed from these as much as possible. Now, this is the domain of remote sensing, which will be the subject of further study in another module, using the textbook *Principles of Remote Sensing* [30].

An important distinction that we must make is that between 'image' and 'raster'. By the first term, we mean a picture with pixels that represent measured local reflectance values in some designated part of the electro-magnetic spectrum. No value has yet been added in terms of interpreting such values as thematic or geographic characteristics. When we use the term 'raster', we assume this value-adding interpretation has been carried out. With an image, we talk of its constituent pixels; with a raster we talk of its cells.

In practice, it is not always feasible to obtain spatial data using these techniques. Factors of cost and available time may be a hindrance, and moreover, previous projects sometimes have acquired data that may fit the current project's purpose. we look at some of the 'indirect' techniques of using existing sources below.

4.1.2 Digitizing paper maps

A cost-effective, though indirect, method of obtaining spatial data is by digitizing existing maps. This can be done through a number of techniques, all of which obtain a digital version of the original (analog) map. Before adopting this approach, one must be aware that, due to the indirect process, positional errors already in the paper map will further accumulate, and that one is willing to accept these errors.

In *manual digitizing*, a human operator follows the map's features (mostly lines) with a mouse device, and thereby traces the lines, storing location coordinates relative to a number of previously defined *control points*. Control points are sometimes also called 'tie points'. Their function is to 'lock' a coordinate system onto the digitized data: the control points on the map have *known* coordinates, and 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 a check on the positional errors made. There are two forms of digitizing: *on-tablet* and *on-screen* manual digitizing.

In on-tablet digitizing, the original map is fitted on a special tablet and the operator moves a special tablet mouse over the map, selecting important points. In on-screen digitizing, a scanned image of the map—or in fact, some other image—is shown on the computer screen, and the operator moves an ordinary mouse cursor over the screen, again selecting important points. In both cases, the GIS works as a point recorder, and from this recorded data, line features are later constructed. There are usually two modes in which the GIS can record: in *point mode*, the system only records a mouse location when the operator says so; in *stream mode*, the system almost continuously records locations. The first is the more useful technique because it can be better controlled, as it is less prone to shaky hand movements.

Another set of techniques also works from a scanned image of the original map, but uses the 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 is 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 on relatively simple sources.

The scanning process

A digital scanner illuminates a to-be-scanned document and measures with a sensor the intensity of the reflected light. The result of the scanning process is an image as a matrix of pixels, each of which holds a reflectance value. Before scanning, one has to decide whether to scan the document in *line art*, *grey-scale* or *colour* mode. The first results in either 'white' or 'black' pixel values; the second in one of 256 'grey' values per pixel, with white and black as extremes. An example of the grey-scale scanning process is illustrated in Figure 4.1, with the original document indicated schematically on the left. For colour mode scanning, more storage space is required as a pixel value is represented in a red-scale value, a green-scale value and a blue-scale value. Each of these three scales, like in the grey-scale case, allows 256 different values.

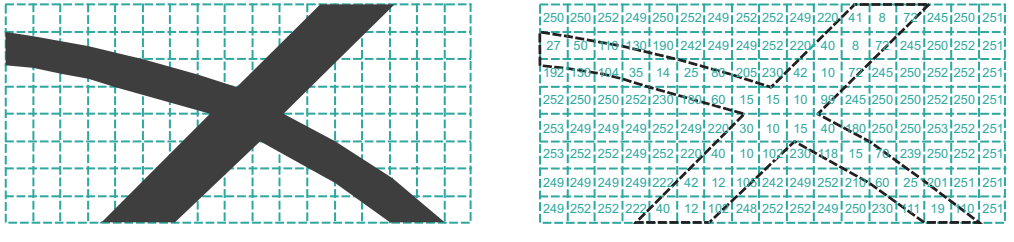


Figure 4.1: The input and output of a (grey-scale) scanning process: (a) the original document in black (with scanner resolution in green), (b) scanned document with grey-scale pixel values (0 = black, 255 = white)

Digital 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). One may opt not to use a scanner at its maximal resolution but at a lower one, depending on the requirements for use. For manual on-screen digitizing of a paper map, a resolution of 200–300 dpi is usually sufficient, depending on

the thickness of the thinnest lines. For manual on-screen digitizing of aerial photographs, higher resolutions are recommended—typically, at least 800 dpi. (Semi-)automatic digitizing requires a resolution that results in scanned lines of at least three pixels wide to enable the computer to trace the centre of the lines and thus avoid displacements. For paper 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 with various techniques of image processing. This may include corrections of colour, brightness and contrast, or the removal of noise, the filling of holes, or the smoothing of lines. It is important to understand that a scanned image is *not* a structured data set of classified and coded objects. Additional, sometimes hard, work is required to associate categories and other thematic attributes with the recognized features.

The vectorization process

Vectorization is the process that attempts to distill points, lines and polygons from a scanned image. As scanned lines may be several pixels wide, they are often first 'thinned', to retain only the centreline. This thinning process is also known as *skeletonizing*, as it removes all pixels that make the line wider than just one pixel. The remaining centreline pixels are converted to series of (x, y) coordinate pairs, which define the found polyline. Afterwards, 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.

Semi-automatic vectorization proceeds by placing the mouse pointer at the start of a line to be vectorized. The system automatically performs line-following with the image as input. At junctions, a default direction is followed, or the operator may indicate the preferred direction.

Pattern recognition methods—like Optical Character Recognition (OCR) for text—can be used for the automatic detection of graphic symbols and text. Once symbols are recognized as image patterns, they can be replaced by symbols in vector format, or better, by attribute data. For example, the numeric values placed on contour lines can be detected automatically to attach elevation values to these vectorized contour lines.

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 4.2.

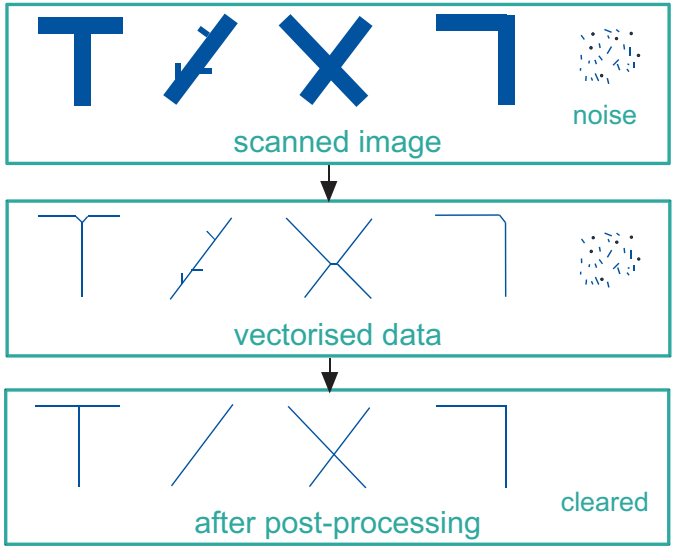


Figure 4.2: The phases of the vectorization process and the various sorts of small error caused by it. The post-processing phase makes the final repairs.

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—like topographic maps and aerial photographs—are therefore better manually digitized. Automatic digitizing in interactive mode is more suitable for images with few types of information that require some interpretation, as is the case in cadastral maps. Fully automatic digitizing is feasible for maps that depict mainly one type of information—as in cadastral boundaries and contour lines. Figure 4.3 provides an overview of these distinctions.

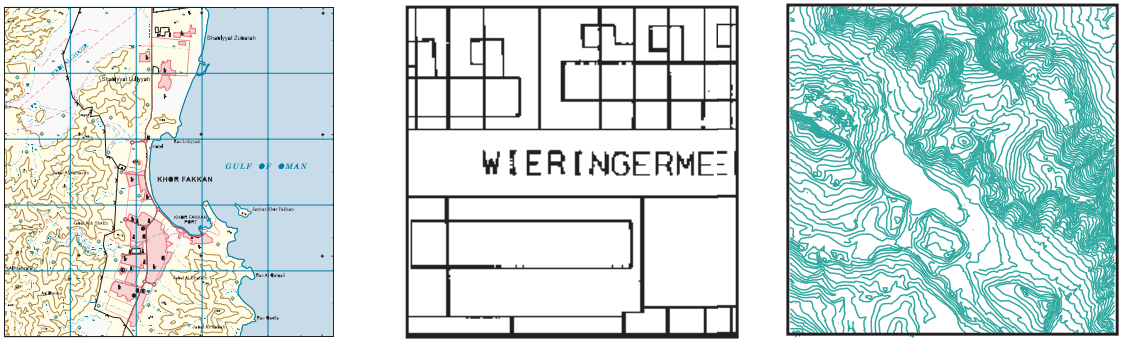


Figure 4.3: The choice of digitizing technique depends on the type of source document.

In practice, when all digitizing techniques are feasible, the optimal one may be a combination of methods. For example, contour line separates can be automatically digitized and used to produce a DEM. Existing topographic maps can

be digitized manually. Geometrically corrected new aerial photographs, with the vector data from the topographic maps displayed on top, can be used for updating by means of manual on-screen digitizing.

4.1.3 Obtaining spatial data elsewhere

Various spatial data sources are available from elsewhere, though sometimes at a price. It all depends on the nature, scale, and date of production that one requires. Topographic base data is easier to obtain than elevation data, which is in turn easier to get than natural resource or census data. Obtaining large-scale data is more problematic than small-scale, of course, while recent data is more difficult to obtain than older data. Some of this data is only available commercially, as usually is satellite imagery.

National mapping organizations (NMOs) historically are the most important spatial data providers, though their role in many parts of the world is changing. Many governments seem to be less willing to maintain large institutes like NMOs, and are looking for alternatives to the nation's spatial data production. Private companies are probably going to enter this market, and for the GIS application people this will mean they no longer have a single provider.

Statistical, thematic data always was the domain of national census or statistics bureaus, but they too are affected by changing policies. Various commercial research institutes also are starting to function as provider for this type of information.

Clearinghouses As digital data provision is an expertise by itself, many of the above-mentioned organizations dispatch their data via centralized places, essentially creating a marketplace where potential data users can 'shop'. It will be no surprise that such markets for digital data have an entrance through the world-wide web. They are sometimes called *spatial data clearinghouses*. The added value that they provide is to-the-point metadata: searchable descriptions of the data sets that are available. We discuss clearinghouses further in Section 7.4.3.

Data formats An important problem in any environment involved in digital data exchange is that of *data formats* and *data standards*. Different formats were implemented by different GIS vendors; different standards come about with different standardization committees.

The good news about both formats and standards is that there are so many to choose from; the bad news is that this causes all sorts of conversion problems. We will skip the technicalities—as they are not interesting, and little can be learnt from them—but warn the reader that conversions from one format to another may mean trouble. The reason is that not all formats can capture the same information, and therefore conversions often mean loss of information. If one obtains a spatial data set in format F, but wants it in format G, for instance because the locally preferred GIS package requires it, then usually a conversion function can be found, likely in that same GIS. The proof of the pudding is to also find an inverse conversion, back from G to F, and to ascertain whether the double conversion back to F results in the same data set as the original. If this is the case, both conversions are not causing information loss, and can safely be applied. More on spatial data format conversions can be found in 7.4.1.

4.2 Spatial referencing

In the early days of GIS, users were handling spatially referenced data from a single country. The data was derived from paper maps published by the country's mapping organization. Nowadays, GIS users are combining spatial data from a certain country with global spatial data sets, reconciling spatial data from a published map with coordinates established with satellite positioning techniques and integrating spatial data from neighbouring countries. To perform these tasks successfully, GIS users need a certain level of appreciation for a few basic spatial referencing concepts pertinent to published maps and spatial data.

Spatial referencing encompasses the definitions, the physical/geometric constructs and the tools required to describe the geometry and motion of objects near and on the Earth's surface. Some of these constructs and tools are usually itemized in the legend of a published map. For instance, a GIS user may encounter the following items in the *map legend* of a conventional published large-scale topographic map: the name of the local vertical datum (e.g., Tide-gauge Amsterdam), the name of the local horizontal datum (e.g., Potsdam Datum), the name of the reference ellipsoid and the fundamental point (e.g., Bessel Ellipsoid and Rauenberg), the type of coordinates associated with the map grid lines (e.g., geographic coordinates, plane coordinates), the map projection (e.g., Universal Transverse Mercator projection), the map scale (e.g., $1 : 25,000$), and the transformation parameters from a global datum to the local horizontal datum.

In the following subsections we shall explain the meaning of these items. An appreciation of basic spatial referencing concepts will help the reader identify potential problems associated with incompatible spatially referenced data.

4.2.1 Spatial reference systems and frames

The geometry and motion of objects in 3D Euclidean space are described in a reference coordinate system. A reference coordinate system is a coordinate system with well-defined origin and orientation of the three orthogonal, coordinate axes. We shall refer to such a system as a *Spatial Reference System* (SRS).

A spatial reference system is a mathematical abstraction. It is realized (or materialized) by means of a *Spatial Reference Frame* (SRF). We may visualize an SRF as a catalogue of coordinates of specific, identifiable point objects, which implicitly materialize the coordinate axes of the SRS. Object geometry can then be described by coordinates with respect to the SRF. An SRF can be made accessible to the user, an SRS cannot. The realization of a spatial reference system is far from trivial. Physical models and assumptions for complex geophysical phenomena are implicit in the realization of a reference system. Fortunately, these technicalities are transparent to the user of a spatial reference frame.

Several spatial reference systems are used in the Earth sciences. The most important one for the GIS community is the *International Terrestrial Reference System* (ITRS). The ITRS has its origin in the centre of mass of the Earth. The Z-axis points towards a mean Earth north pole. The X-axis is oriented towards a mean Greenwich meridian and is orthogonal to the Z-axis. The Y-axis completes the right-handed reference coordinate system (Figure 4.4(a)).

The ITRS is realized through the *International Terrestrial Reference Frame* (ITRF), a catalogue of estimated coordinates (and velocities) at a particular epoch of several specific, identifiable points (or stations). These stations are more or less homogeneously distributed over the Earth surface. They can be thought of as defining the vertices of a *fundamental polyhedron*, a geometric abstraction of the

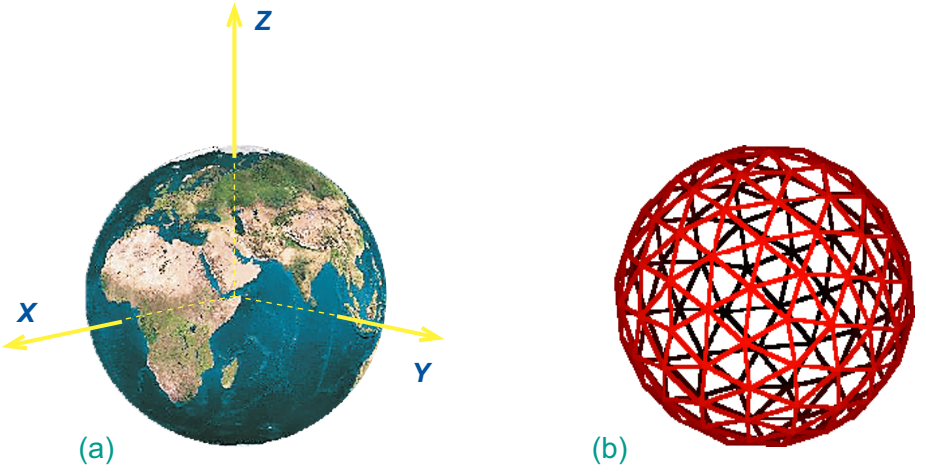


Figure 4.4: (a) The International Terrestrial Reference System (ITRS), and (b) the International Terrestrial Reference Frame (ITRF) visualized as the fundamental polyhedron. Data source for (b): Martin Trump, United Kingdom.

Earth's shape at the fundamental *epoch*[1] (Figure 4.4(b)). Maintenance of the spatial reference frame means relating the rotated, translated and deformed polyhedron at a later epoch to the fundamental polyhedron. Frame maintenance is necessary because of geophysical processes (mainly tectonic plate motion) that deform the Earth's crust at *measurable* global, regional and local scales. The ITRF is ideally suited to describe the geometry and behaviour of moving and stationary objects on and near the surface of the Earth.

Global, geocentric spatial reference systems, such as the ITRS, became available only recently with advances in extra-terrestrial positioning techniques.[2] Since

[1] For the purposes of this book, an epoch is a specific calendar date.

[2] Extra-terrestrial positioning techniques include Satellite Laser Ranging (SLR), Lunar Laser Ranging (LLR), Global Positioning System (GPS), Very Long Baseline Interferometry (VLBI) *et cetera*.

the centre of mass of the Earth is directly related to the size and shape of satellite orbits (in the case of an idealized spherical Earth it is one of the focal points of the elliptical orbits), observing a satellite (natural or artificial) can pinpoint the centre of mass of the Earth, and hence the origin of the ITRS. Before the space age—roughly before the 1960s—it was impossible to realize geocentric reference systems at the accuracy level required for large-scale mapping.

If the ITRF is implemented in a region in a modern way, GIS applications can be conceived that were unthinkable before. Such applications allow for real time spatial referencing and real time production of spatial information, and include electronic charts and electronic maps, precision agriculture, fleet management, vehicle dispatching and disaster management. What do we mean by a 'modern implementation' of the ITRF in a region? First, a regional densification of the ITRF polyhedron through additional vertices to ensure that there are a few coordinated reference points in the region under consideration. Secondly, the installation at these coordinated points of permanently operating satellite positioning equipment (i.e., GPS receivers and auxiliary equipment) and communication links. Examples for (networks consisting of) such permanent tracking stations are the AGRS in the Netherlands and the SAPOS in Germany (refer for both to Appendix A).

The ITRF continuously evolves as new stations are added to the fundamental polyhedron. As a result, we have different realisations of the same ITRS, hence different ITRFs. A specific ITRF is therefore codified by a year code. One example is the ITRF96. ITRF96 is a list of geocentric coordinates (X, Y and Z in metres) and velocities ($\delta X/\delta t$, $\delta Y/\delta t$ and $\delta Z/\delta t$ in metres per year) for all stations, together with error estimates. The station coordinates relate to the epoch 1996.0. To obtain the coordinates of a station at any other time (e.g., for epoch 2000.0) the station velocity has to be applied appropriately.

4.2.2 Spatial reference surfaces and datums

It would appear that a specific International Terrestrial Reference Frame is sufficient for describing the geometry and behaviour in time of objects of interest near and on the Earth surface in terms of a uniform triad of geocentric, Cartesian X, Y, Z coordinates and velocities. Why then do we need to also introduce spatial reference surfaces?

Splitting the description of 3D location in 2D (horizontal[3]) and 1D (height) has a long tradition in Earth sciences. With the overwhelming majority of our activities taking place on the Earth's topography, a complex 2D curved surface, we humans are essentially inhabitants of 2D space. In first instance, we have sought intuitively to describe our environment in two dimensions. Hence, we need a *simple* 2D curved reference surface upon which the complex 2D Earth topography can be projected for easier 2D horizontal referencing and computations. We humans, also consider height an add-on coordinate and charge it with a physical meaning. We state that point A lies higher than point B, if water can flow from A to B. Hence, it would be ideal if this simple 2D curved reference surface could *also* serve as a reference surface for heights with a physical meaning.

[3]Caution: horizontal does not mean flat.

The geoid and the vertical datum

To describe heights, we need an imaginary surface of zero height. This surface must also have a physical meaning, otherwise it cannot be sensed with instruments. A surface where water does not flow, a level surface, is a good candidate. Any sensor equipped with a bubble can sense it. Each level surface is a surface of constant height. However, there are infinitely many level surfaces. Which one should we choose as *the* height reference surface? The most obvious choice is the level surface that most closely approximates all the Earth's oceans. We call this surface *the geoid*. Every point on the geoid has the same zero height all over the world. This makes it an ideal global reference surface for heights. How is the geoid realized on the Earth surface in order to allow height measurements?

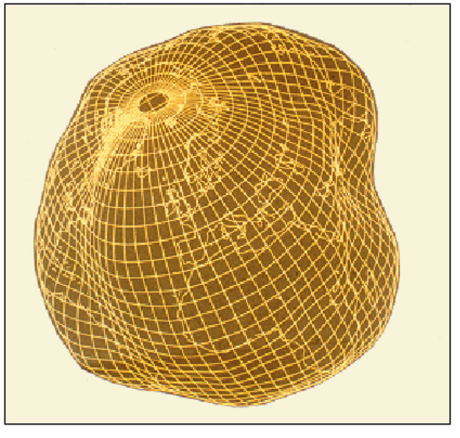


Figure 4.5: The geoid, exaggerated to illustrate the complexity of its surface. Source: Denise Dettmering, *Seminar Notes for Bosch Telekom*, Stuttgart, 2000.

Historically, the geoid has been realized only locally, not globally. A local mean sea level surface is adopted as the zero height surface of the locality. How can the *mean sea level* value be recorded locally? Through the readings, averaged

over a sufficient period of time, of an automatically recording tide-gauge placed in the water at the desired location. For the Netherlands and Germany, the local mean sea level is realized through 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*. The result of this process will be the height above local mean sea level for the Enschede point.

Obviously, there are several realizations of local mean sea levels, also called *local vertical datums*, in the world. They are parallel to the geoid but offset by up to a couple of metres. This offset is due to local phenomena such as ocean currents, tides, coastal winds, water temperature and salinity at the location of the tide-gauge.

The local vertical datum is implemented through a *levelling network*. A levelling network consists of benchmarks, whose height above mean sea level has been determined through geodetic levelling. The implementation of the datum enables easy user access. The users do not need to start from scratch (i.e., from the Amsterdam tide-gauge) every time they need to determine the height of a new point. They can use the benchmark of the levelling network that is closest to the point of interest.

The ellipsoid and the horizontal datum

We have defined a physical construct, the geoid, that can serve as a reference surface for heights. We have also seen how a local version thereof, the local mean sea level, can be realized. Can we also use the local mean sea level surface to project upon it the rugged Earth topography? In principle yes, but in practice no. The mean sea level is everywhere orthogonal to the direction of the gravity vector. A surface that must satisfy this condition is bumpy and complex to describe mathematically. It is rather difficult to determine 2D coordinates on this surface and to project this surface onto a flat map. Which mathematical reference surface is then more appropriate? The mathematical shape that is simple enough and most closely approximates the local mean sea level is the surface of an *oblate ellipsoid*. How is this mathematical surface realized?

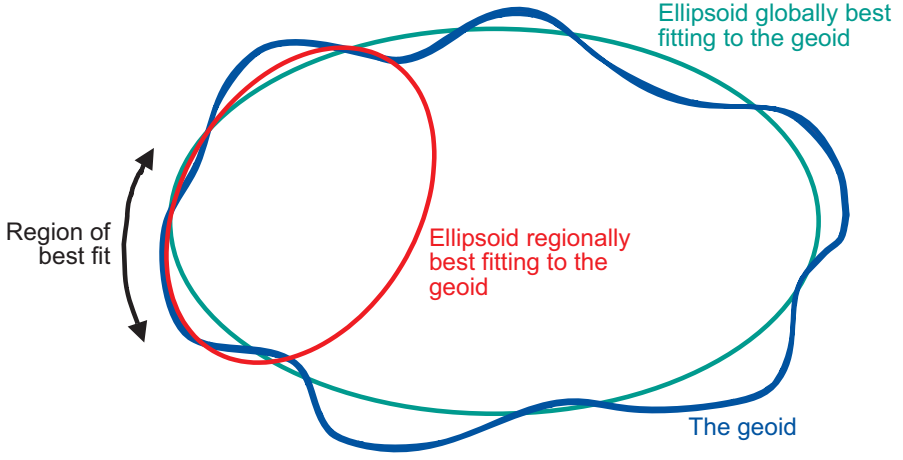


Figure 4.6: The geoid, a globally best fitting ellipsoid for it, and a regionally best fitting ellipsoid for it, for a chosen region. Adapted from: Ordnance Survey of Great Britain. *A Guide to Coordinate Systems in Great Britain*, see Appendix A.

Historically, the ellipsoidal surface has been realized locally, not globally. An

ellipsoid with specific dimensions—a and b as half the length of the major, respectively minor, axis—is chosen which best fits the local mean sea level. Then the ellipsoid is positioned and oriented with respect to the local mean sea level by adopting a latitude (ϕ) and longitude (λ) and height (h) of a so-called fundamental point and an azimuth to an additional point. We say that a *local horizontal datum* is defined by

(a) dimensions (a, b) of the ellipsoid,

(b) the adopted geographic coordinates ϕ and λ and h of the fundamental point, and

(c) azimuth from this point to another.

There are a few hundred local horizontal datums in the world. The reason is obvious. Different ellipsoids with varying position and orientation had to be adopted to best fit the local mean sea level in different countries or regions (Figure 4.6).

An example is the Potsdam datum, the local horizontal datum used in Germany. The fundamental point is 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 the new point in the Potsdam datum.

The local horizontal datum is implemented through a so-called *triangulation network*. A triangulation network consists of monumented points forming a network of triangular mesh elements. The angles in each triangle are measured in addition to at least one side of a 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. The implementation of the datum enables easy user access. The users do not need to start from scratch (i.e., from the fundamental point Rauenberg) 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.

Local and global datums

We described the need for defining additional reference surfaces and introduced two constructs, the local mean sea level and the ellipsoid. We saw how they can be realized as vertical and horizontal datums. We mentioned how they can be implemented for height and horizontal referencing. Most importantly, we saw that realizations of these surfaces are made locally and have resulted in hundreds of local vertical and horizontal datums worldwide. Are a *global vertical datum* and a *global horizontal datum* possible?

The good news is that a geocentric ellipsoid, known as the *Geodetic Reference System 1980 (GRS80) ellipsoid* (refer to Appendix A, GRS80), can now be realized thanks to advances in extraterrestrial positioning techniques. The global horizontal datum is a realization of the GRS80 ellipsoid. The trend is to use the global horizontal datum everywhere in the world for reasons of global compatibility. The same will soon hold true for the geoid as well. Launches for gravity satellite missions are planned in the next few years by the American and European space agencies. These missions will render an accurate global geoid. Why are we looking forward to an accurate global geoid?

We are now capable of determining a triad of Cartesian (X, Y, Z) geocentric coordinates of a point with respect to the ITRF with an accuracy of a few centimetres. We can easily transform this Cartesian triad into geographic coordinates (ϕ, λ, h) with respect to the geocentric, global horizontal datum without loss of accuracy. However, the height h, obtained through this straightforward transformation, is devoid of physical meaning and contrary to our intuitive human perception of a height. Moreover, height H, above the geoid is currently two orders of magnitude less accurate. The satellite gravity missions, will allow the determination of height H, above the geoid with centimetre level accuracy for the first time. It is foreseeable that global 3D spatial referencing, in terms of

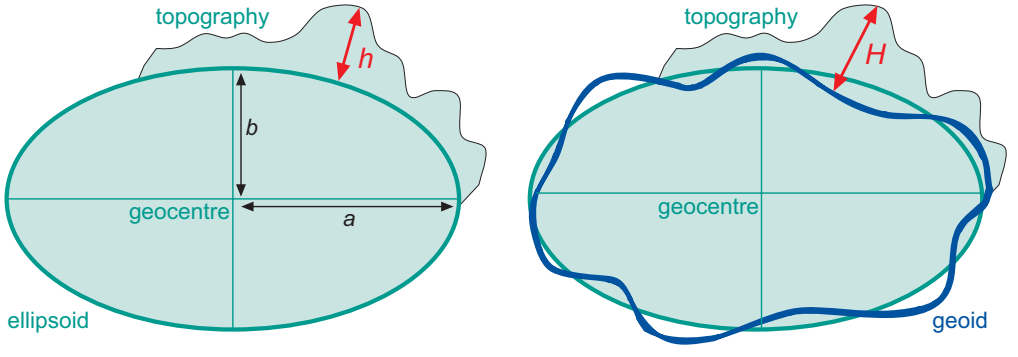


Figure 4.7: Height h above the geocentric ellipsoid, and height H above the geoid. The first is measured orthogonal to the ellipsoid, the second orthogonal to the geoid.

(ϕ, λ, H), shall become ubiquitous in the next 10–15 years. If all published maps are also globally referenced by that time, the underlying spatial referencing concepts will become transparent and irrelevant to GIS users.

The bad news is that the hundreds of existing local horizontal and vertical datums are still relevant because they are implicit in map products all over the world. For the next several years, we shall be dealing with both local and global datums until the former are eventually phased out. During the transition period, we shall need tools to transform coordinates from local horizontal datums to a global horizontal datum and *vice versa*. The organizations that usually develop transformation tools and make them available to the user community are provincial or national mapping organizations and cadastral authorities.

4.2.3 Datum transformations

The rationale for adopting a global geocentric datum is the need for compliance with international best practice and standards [49] (and refer to Appendix A, LINZ). Satellite positioning and navigation technology, now widely used around the world for spatial referencing, implies a global geocentric datum. Also, the complexity of spatial data processing relies heavily on software packages that are designed for, and sold to, global markets. As more countries go global the cost of being different (in our case, the cost of maintaining a local datum) will increase. Finally, global and regional data sets (e.g., for global environmental monitoring) refer nowadays almost always to a global geocentric datum and are useful to individual nations only if they can be reconciled with the local datum.

How do mapping organizations react to this challenge? Let us take a closer look at a typical reaction. The Land Information New Zealand (LINZ) recently adopted the International Terrestrial Reference System (ITRS) and a geocentric horizontal datum, based on the GRS80 ellipsoid. The ITRS will be materialized in New Zealand through ITRF96 at epoch 2000.0 [38]. LINZ has launched an intensive publicity campaign to help its customers get in step with the new geocentric datum [29]. LINZ advises the user community to develop and implement strategies to cope with the change and proposes different approaches (e.g., change all at once, change by product/region, change upon demand). They also advise the users to audit existing data and sources, to establish procedures for converting to the new datum and for dealing with dual coordinates during the transition, and to adopt procedures for changing legislation.

Mapping organizations do not only coach the user community about the implications of the geocentric datum. They also develop tools to enable users to transform coordinates of spatial objects from the new datum to the old one. This process is known as *datum transformation*. The tools are called *datum transforma-*

tion parameters. Why do the users need these transformation parameters? Because, they are typically collecting spatial data in the field using satellite navigation technology. They also typically need to represent this data on a published map based on a local horizontal datum.

The good news is that a transformation from datum A to datum B is a mathematically straightforward process. Essentially, it is a transformation between two orthogonal Cartesian spatial reference frames together with some elementary tools from adjustment theory. In 3D, the transformation is expressed with seven parameters: three rotation angles (α, β, γ), three origin shifts (X_0, Y_0, Z_0) and one scale factor s. The input in the process are coordinates of points in datum A and coordinates of the same points in datum B. The output is an estimate of the transformation parameters and a measure of the likely error of the estimate.

The bad news is that the estimated parameters may be inaccurate if the coordinates of the common points are wrong. This is often the case when we transform coordinates from a local horizontal datum to a 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 estimate will depend on which particular common points are chosen, and they also will depend on whether all seven parameters, or only a sub-set of them, are estimated.

Here is an illustration of what we may expect. The example below is concerned with 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:

- The set provided by the federal mapping organization (labelled 'National set' in Table 4.1) was calculated using common points distributed throughout Germany. This set contains all seven parameters and is valid for all of Germany.
- The set provided by the mapping organization of Baden-Württemberg (labelled 'Provincial set' in Table 4.1) has been calculated using common points distributed throughout the province of Baden-Württemberg. This set contains all seven parameters and is valid only within the borders of that province.
- The set provided by the National Imagery and Mapping Agency (NIMA) of the USA (labelled 'NIMA set' in Table 4.1) has been calculated using common points distributed throughout Germany. This set contains a coordinate shift only (no rotations, and scale equals unity). It is valid for all of Germany.

	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	X_0	-581.99 m	-518.19 m	-635 m
	Y_0	-105.01 m	-43.58 m	-27 m
	Z_0	-414.00 m	-466.14 m	-450 m

Table 4.1: Transformation of Cartesian coordinates; this 3D transformation provides seven parameters, scale factor s, the rotation angles α, β, γ, and the origin shifts X_0, Y_0, Z_0.

The three sets of transformation parameters vary by several tens of metres, for the aforementioned reasons. These sets of transformation parameters have

been used to transform the ITRF cartesian coordinates of a point in the state of Baden-Württemberg. The ITRF (X, Y, Z) coordinates are

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

The three sets of transformed coordinates in the Potsdam datum are:

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

It is obvious that the three sets of transformed coordinates agree at the level of a few metres. In a different country, the agreement could be at the level of centimetres, or tens of metres and this depends primarily on the quality of implementation of the local horizontal datum. It is advisable that GIS users act with caution when dealing with datum transformations and that they consult with their national mapping organization, wherever appropriate (refer to Appendix A, Ordnance Survey).

4.2.4 Map projections

To represent parts of the surface of the Earth on a flat paper map or on a computer screen, the curved horizontal reference surface must be mapped onto the 2D mapping plane. The reference surface is usually an oblate ellipsoid for large-scale mapping, and a sphere for small-scale mapping.[4] Mapping onto a 2D mapping plane means assigning plane Cartesian coordinates (x, y) to each point on the reference surface with geographic coordinates (ϕ, λ), see Figure 4.8.

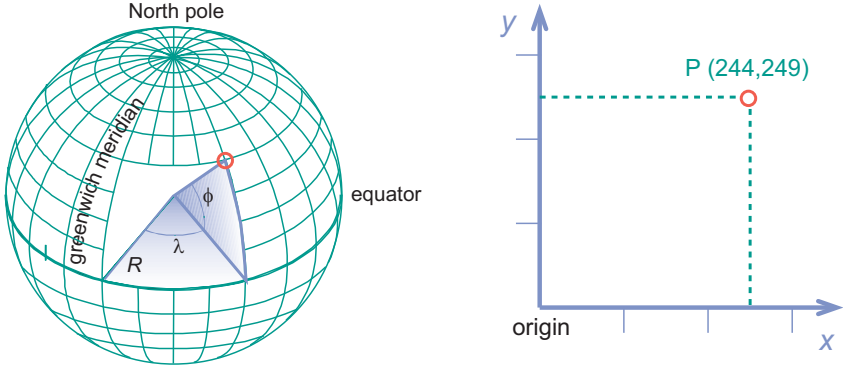


Figure 4.8: Two 2D spatial referencing approaches: (a) through geographic coordinates (ϕ, λ); (b) through Cartesian plane, rectangular coordinates (x, y).

[4] In practice, maps at scale $1 : 1,000,000$ or smaller can use the mathematically simpler sphere without the risk of large distortions.

Classification of map projections

Any map projection is associated with distortions. There is simply no way to flatten out a piece of ellipsoidal or spherical surface without stretching some parts of the surface more than others. Some map projections can be visualized as true *geometric projections* directly onto the mapping plane, or onto an intermediate surface, which is then rolled out into the mapping plane. Typical choices for such intermediate surfaces are cones and cylinders. Such map projections are then called *azimuthal*, *conical*, and *cylindrical*, respectively. Figure 4.9 shows the surfaces involved in these three classes of projections.

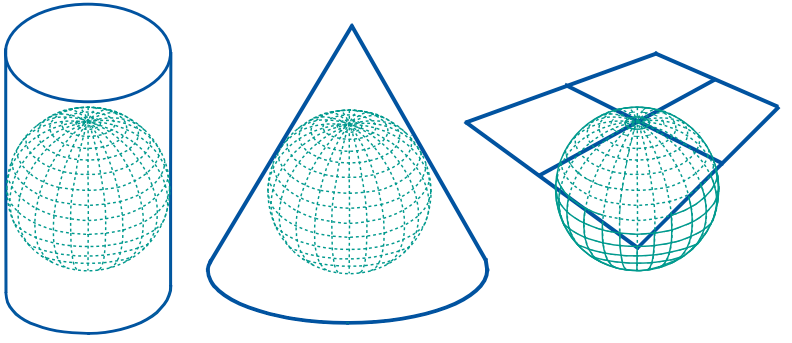


Figure 4.9: Classes of map projections

The planar, conical, and cylindrical surfaces in Figure 4.9 are all *tangent* surfaces; they touch the horizontal reference surface in one point (plane) 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; illustrations are in Figure 4.10. Then, the reference surface is intersected

along one closed line (plane) or two closed lines (cone and cylinder).

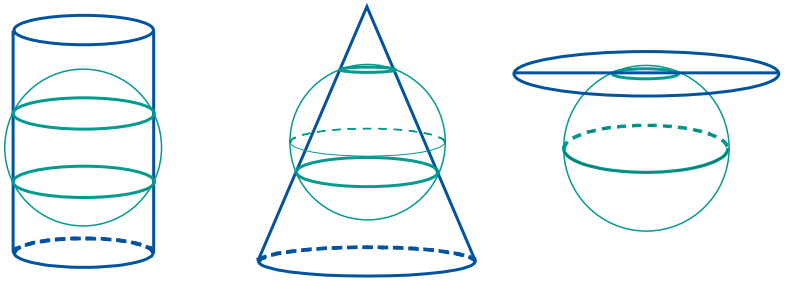


Figure 4.10: Three secant projection classes

In the geometrical depiction of map projections in Figure 4.9 and 4.10, the symmetry axes of the plane, cone and cylinder coincide with the rotation axis of the ellipsoid or sphere. In this case, the projection is said to be a *normal projection*. The other cases are *transverse projection* (symmetry axis in the equator) and *oblique projection* (symmetry axis is somewhere between the rotation axis and equator of the ellipsoid or sphere). These cases are illustrated in Figure 4.11.

So far, we have not specified *how* the curved horizontal reference surface is projected onto the plane, cone or cylinder. This *how* determines which 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:

- In a conformal map projection the angles between lines on the curved reference surface are identical to the angles between the images of these lines in the map.

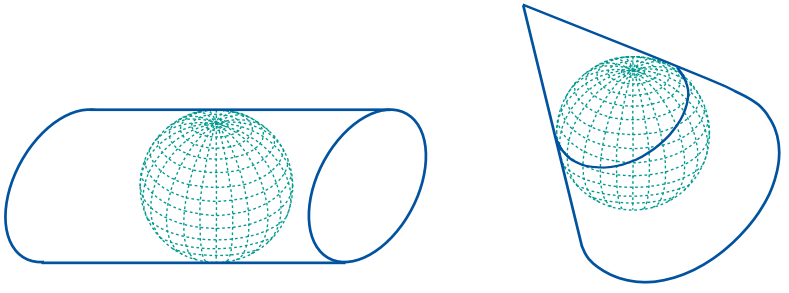


Figure 4.11: A transverse and an oblique projection

- In an equal-area (equivalent) map projection the area enclosed by the lines in the map is representative of—modulo the map scale—the area enclosed by the original lines on the curved reference surface.
- In an equidistant map projection the length of particular lines in the map is representative of—modulo the map scale—the length of the original lines on the curved reference surface.

A particular map projection can have any one of these three properties. Conformality and equivalence are mutually exclusive.

Based on these discussions, a particular map projection can be classified. An example would be the classification 'conformal conic projection with two standard parallels' having the meaning, 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; i.e., the cone is secant and the cone's symmetry axis is parallel to the rotation axis.

Often, a particular type of map projection is also named after its inventor (or

first publisher). For example, the ‘conformal conic projection with two standard parallels’ is also referred to as ‘Lambert’s conical projection’ [27].

Mapping equations

The actual mapping is not done through the aforementioned geometric projections, but through mapping equations. (Some of the mapping equations in use cannot be visualized as a geometric projection.) A *forward mapping equation* associates mathematically the plane Cartesian coordinates (x, y) of a point to the geographic coordinates (ϕ, λ) of the same point on the curved reference surface:

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

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

$$(\phi, \lambda) = f^{-1}(x, y).$$

Equations like these can be specified for all of the map projections discussed in the previous section. More importantly, they can also be specified for a number of map 'projections' that do not have the kind of geometric interpretation as discussed above, e.g., the so-called Gauss-Krüger projection.

Change of map projection

Forward and inverse mapping equations are normally used to transform data from one map projection to 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) to 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. The principles are illustrated in Figure 4.12.

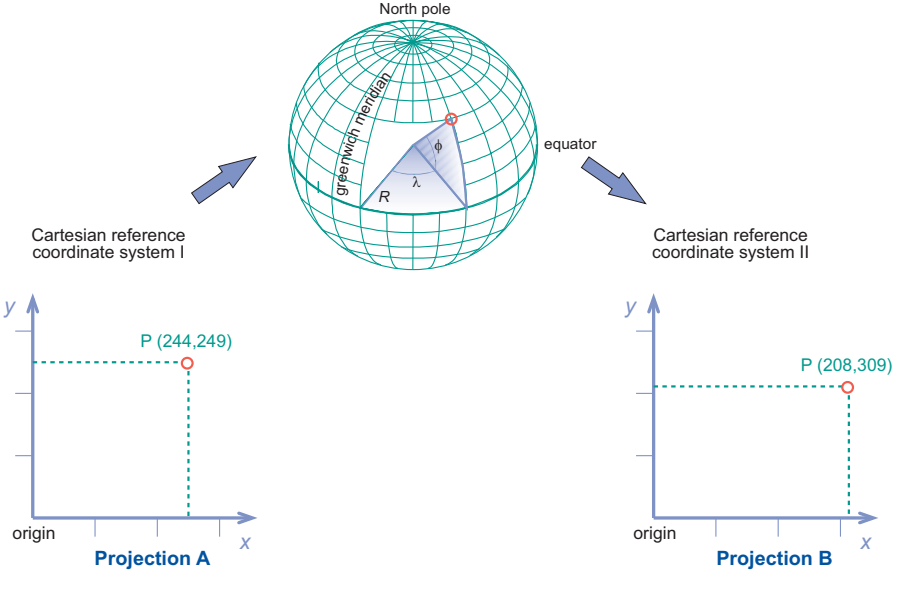


Figure 4.12: The principle of changing from one into another map projection

Historically, a GIS has handled data referenced spatially with respect to the (x, y) coordinates of a specific map projection. For GIS application domains requiring 3D spatial referencing, a height coordinate may be added to the (x, y) coordinate of the point. The additional height coordinate can be a height H above mean sea level, which is a height with a physical meaning. These (x, y, H) coordinates can be used to represent objects in a 3D GIS.

4.3 Data preparation

Spatial data preparation aims to make the 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 overshoots of lines at intersections, deleting duplicate lines, closing gaps in lines, and generating polygons. Data may need to be converted to either vector format or raster format to match other data sets. Additionally, the process includes associating attribute data with the spatial data through either manual input or reading digital attribute files into the GIS/DBMS.

The intended use of the acquired spatial data, furthermore, may require to thin the data set and retain only the features needed. The reason may be that not all features are relevant for subsequent analysis or subsequent map production. In these cases, data and/or cartographic generalization must be performed to restrict the original data set.

4.3.1 Data checks and repairs

Acquired data sets must be checked for consistency and completeness. This requirement applies to the *geometric* and *topological* quality as well as the *semantic* quality of the data.

There are different approaches to clean up data. Errors can be identified automatically, after which manual editing methods can be applied to correct the errors. Alternatively, a system may identify and automatically correct many errors. 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. A number of clean-up operations is illustrated in Table 4.2.

With polygon data, one usually starts with many polylines that are combined in the first step (from Figure 4.13(a) to (b)). This results in fewer polylines (with more internal vertices). Then, polygons can be identified (c). Sometimes, polylines do not connect to form closed boundaries, and therefore must be connected; this step is not indicated in the figure. In a final step, the elementary topology of the polygons can be deduced (d).

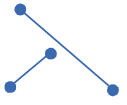
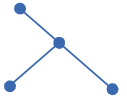
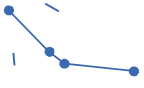
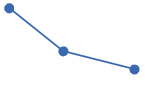
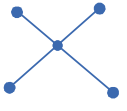
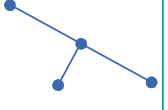
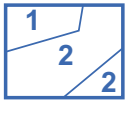
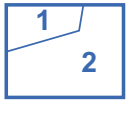
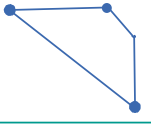
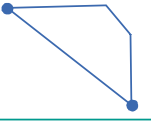
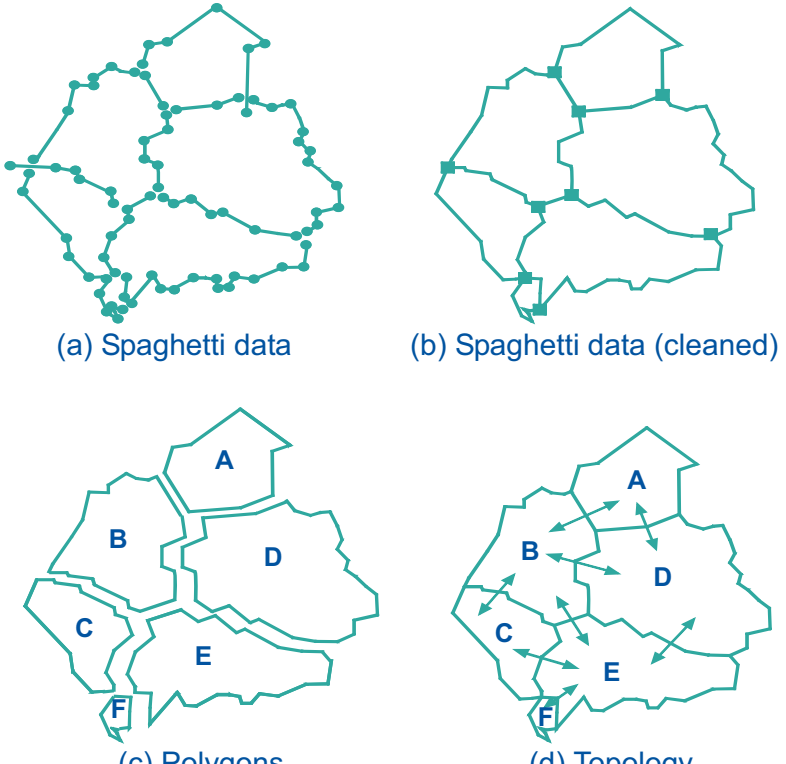
Before cleanup	After cleanup	Description	Before cleanup	After cleanup	Description
		Erase duplicates or sliver lines			Extend undershoots
		Erase short objects			Snap clustered nodes
		Break crossing objects			Erase dangling objects or overshoots
		Dissolve polygons			Dissolve nodes into vertices

Table 4.2: The first clean-up operations for vector data



(a) Spaghetti data (b) Spaghetti data (cleaned)

(c) Polygons (d) Topology

Figure 4.13: Continued clean-up operations for vector data, turning spaghetti data into topological structure.

Associating attributes

Attributes may be automatically associated with features, when they have been given unique identifiers. We discussed such techniques already in Section 3.3.6.

In vector data, attributes are assigned directly to the features, while in a raster the attributes are assigned to all cells that represent a feature.

Rasterization or vectorization

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 is known as *rasterization*. It 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 too large cell size 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 the raster resolution is too small, the raster will easily become too big.

Rasterization somehow is a step backward: raster cell conglomerates of which the boundary is only an approximation of the objects' original boundary replace objects for which accurate geometrical representation was available. The reason to perform it nonetheless lies in the integrated use later with some other data source that we only have as raster, and cannot vectorize (easily).

An alternative way to rasterization is to not perform it during the data preparation phase, but to use GIS rasterization functions on-the-fly, that is when the computations call for it. This allows keeping the vector data and generating raster data from them when needed. Obviously, the issue of performance trade-off must be looked into. We do not advocate to necessarily work in a purely vector or purely raster setting.

There is an inverse operation, called *vectorization*, that produces a vector data set from a raster. We have looked at this in some sense already: namely in the production of a vector set from a scanned image. Another form of vectorization takes place when we want to identify features or patterns in remotely sensed imagery. The keywords here are *feature extraction* and *pattern recognition*, but

these subjects will be dealt with in *Principles of Remote Sensing* [30].

Topology generation

We have already mentioned the identification of polygons from vectorized data sources. More topological relations may sometimes be needed. Examples are the questions of what is connected to what (for instance, in networks), what is the direction of the network's constituent lines, and which lines have over- and underpasses. 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). Many of these questions are mostly questions of data semantics, and can therefore usually only be answered by a human operator.

4.3.2 Combining multiple data sources

A GIS project usually involves multiple data sets, so a next step addresses the issue of how these multiple sets relate to each other. There are three fundamental cases to be considered if we compare data sets pairwise:

- they may be about the same area, but differ in accuracy,
- they may be about the same area, but differ in choice of representation, and
- they may be about adjacent areas, and have to be merged into a single data set.

We look at these situations below. They are best understood with an example.

Differences in accuracy

Images come at a certain resolution, and paper 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 to improve 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 at different resolutions.

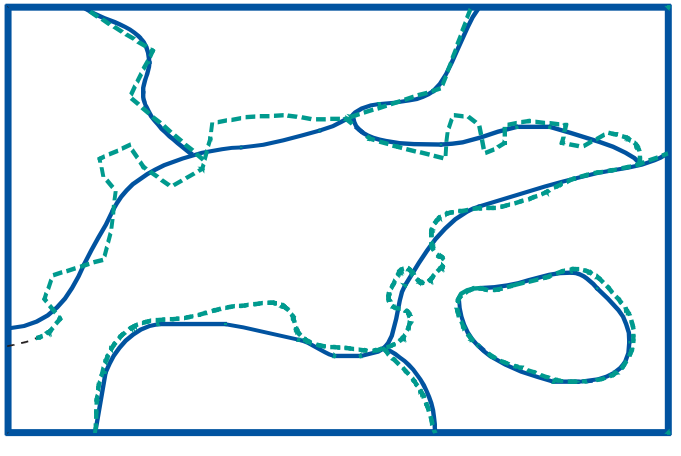


Figure 4.14: The integration of two vector data sets may lead to slivers

In Figure 4.14, the polygons of two digitized maps at different scales are overlaid. Due 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 known as *sliver polygons*. If the map scales involved 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 the issues.

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 from the foundation data a separate database. Such a derived database may be called a *cartographic* database as 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 ($1:5,000 - 1:10,000$) and other databases for the smaller scale ranges. They maintain a *multi-scale* data environment.

Differences in representation

There exist more advanced GIS applications that require the possibility of representing the same geographic phenomenon in different ways. Map production at various map scales is again an example but there are numerous others. The commonality is that phenomena must sometimes be viewed as points, and at other times as polygons, for instance. The complexity that this requirement entails is that the GIS or the DBMS must keep track of links between different representations for the same phenomenon, and must also provide support for decisions as to which representations to use in which situation.

For example, a small-scale national road network analysis may represent villages as point objects, but a nation-wide urban population density study should regard all municipalities as represented by polygons.

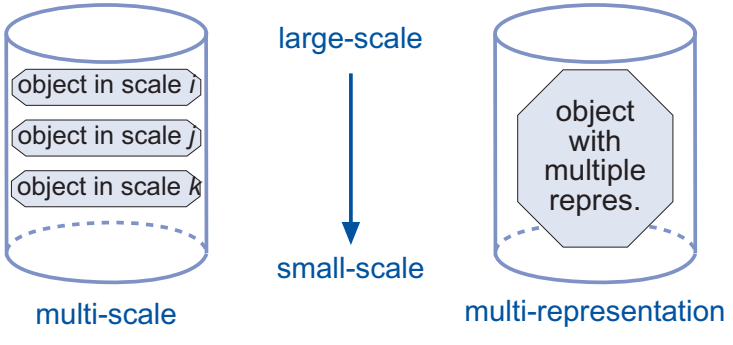


Figure 4.15: Multi-scale and multi-representation systems compared; the main difference is that multi-representation systems have a built-in 'understanding' that different representations belong together.

The links between various representations for the same things maintained by the system allows interactive traversal, and many fancy applications of their use seem possible. The systems that support this type of data traversal are called *multi-representation* systems. A comparison is illustrated in Figure 4.15.

Merging data sets of adjacent areas

When individual data sets have been prepared as described above, they sometimes have to be matched together such that a single 'seamless' data set results, and that the appearance of the integrated geometry is as homogeneous as possible.

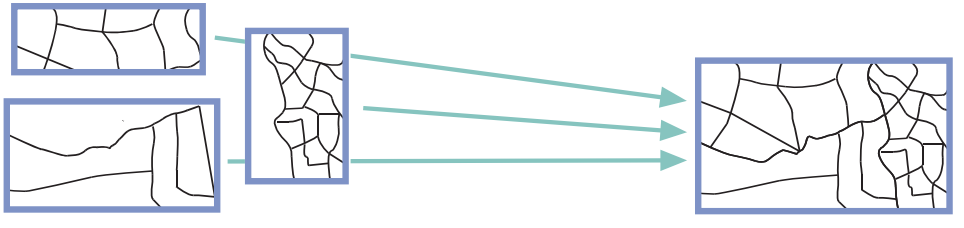


Figure 4.16: Multiple adjacent data sets, after cleaning, can be matched and merged into a single one.

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 4.16 illustrates a typical situation.

Some GISs have merge or edge-matching functions to solve the problem arising from merging adjacent data. *Edge-matching* is an editing procedure used to ensure that all features along shared borders have the same edge locations. Coordinates of the objects along shared borders are adjusted to match those in the neighbouring data sets. Mismatches may still be possible, so a visual check, and interactive editing is likely to be needed.

4.4 Point data transformation

A common situation—particularly, but not only, in the Earth sciences—is that one of the subjects of study is a geographic field. Remember that by our definition, a geographic field associates a value with *each* location in the study area. Clearly, ground-based field surveys cannot possibly obtain measurements for all locations, and only finitely many samples can be taken. Still, ground-based surveys in many cases produce data of a quality that is superior to that of remotely sensed imagery. So, this presents a problem: we want to know (a representation of) the geographic field, but can only take finitely many measurements of it. In GIS data terms, we want to construct a field representation—either as a raster, or as a vector data set—from a point data set. This common problem is the topic of this section.

A fundamental issue is what sort of field we are considering: is it a discrete field—providing geological units, for instance—in which the values are of a qualitative nature, or is it a continuous field—elevation, temperature, salinity *et cetera*—in which the values are of a quantitative nature? This distinction matters, because qualitative data cannot be interpolated, whereas quantitative data can.

A simplistic but hopefully clarifying example is given in Figure 4.17. Our field survey has taken only two measurements, one in P and one in Q. The values obtained in these two locations are represented by a dark and light green tint, respectively. If the field is considered a qualitative field, and we have no further knowledge, the only assumption we can make for other locations is that those nearer to P probably have P's value, whereas those nearer to Q have Q's value. This is illustrated in part (a).

If, on the contrary, our field is considered to be quantitative, meaning that

Figure 4.17: A geographic field representation obtained from two point measurements: (a) for qualitative (categorical), and (b) for quantitative (interpolatable) point measurements. The value measured at P is represented as dark green, that at Q as light green.

we can interpolate values, we can let the values of P and Q contribute both to values for other locations. This is done in part (b) of the figure. To what extent the measurements contribute is determined by the interpolation function. In the figure, the contribution is expressed in terms of the ratio of distances to P and Q. We will see in the sequel that the choice of interpolation function is a crucial factor in any method of field construction from point measurements.

How we represent a field constructed from point measurements in the GIS also depends on the above distinction. A *qualitative (discrete) field* can either be represented as a classified raster or as a polygon data layer, in which each polygon has been assigned a (constant) field value. A *quantitative (continuous) field* can be represented as an unclassified raster, as an isoline (thus, vector) data layer, or perhaps as a TIN. Which option to pick depends (again) on what one wants to do with the data afterwards, during spatial data analysis.

4.4.1 Generating discrete field representations from point data

If the field we want to construct is assumed to be discrete, we cannot interpolate the point measurements. We are thus in the situation of Figure 4.17(a), but obviously with many more point measurements. The best we can do, if we want to have it done automatically by the GIS, is to assume that any location is assigned the value of the closest measured point. Effectively, such a technique will construct areas around the points of measurement that will all be assigned the (categorical) value of the point inside.

Thinking in vector terms, this will mean the construction of Thiessen polygons around the points of measurement. (The boundaries of such polygons, by the way, are the locations for which more than one point of measurement is the closest point.) An illustration is provided is Figure 4.18. More about Thiessen polygons will be discussed in Section 5.4.1.

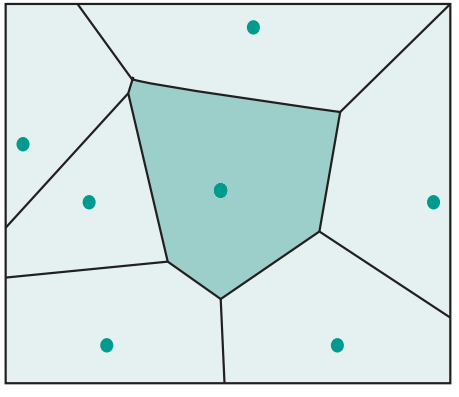


Figure 4.18: Generation of Thiessen polygons for qualitative point measurements. The measured points are indicated in dark green; the darker area indicates all locations assigned with the measurement value of the central point.

If we have a vector data layer with Thiessen polygons, we have assigned the values, and we want to continue operating in vector mode later, then we are

ready here. If we, however, want to continue operating in raster mode later, we must still go through a rasterization procedure of the Thiessen polygons. We discussed this in Section 4.3.1.

Expert knowledge may sometimes be available to assist in obtaining a more realistic discrete field representation. For instance, for a field of geological units, one may know that a zone adjacent to a river in the study area is all sedimentary. For this very reason, one may not have sampled the riverine zone. In such a case, it is probably wise to include in the Thiessen polygon generation extra (fake) measurement points for this riverine zone.

4.4.2 Generating continuous field representations from point data

Things become much more interesting, but also much more complicated, if the field that we want to represent is considered to be continuous. We are now in the situation of Figure 4.17(b), but, again, usually with many more point measurements.

As the field is considered to be continuous, we are allowed to use measured values for *interpolation*. There are many continuous geographic fields—elevation, temperature, ground water salinity are just a few examples. We again would like to use measurements to obtain a GIS representation for the entire field. We discuss two techniques to do so: trend surface fitting and moving window averaging.

Commonly, continuous fields are represented in rasters, and we will almost by default assume that they are. Alternatives exist though, as we have seen in discussions in Chapter 2. The most prominent alternative for continuous field representation is a polyline vector layer, in which the lines are isolines. We will shortly address these issues of representation also.

Trend surface fitting

In trend surface fitting, the assumption is that the entire (continuous) geographic field can be represented by a formula $f(x, y)$ that for given location with coordinates (x, y) will give us the approximated value of the field in that location.

The key quest in trend surface fitting thus is to find out what is the formula that best describes the field. Various classes of formulæ exist, with the simplest being the one that describes a flat, but tilted plane:

$$f(x, y) = c_1 \cdot x + c_2 \cdot y + c_3.$$

If we believe—and this judgement must be based on domain expertise—that the field under consideration can be best approximated by a tilted plane, then the problem of finding the best plane is the problem of determining best values for the coefficients c_1, c_2 and c_3. This is where the point measurements earlier obtained become important. Mathematical techniques, known as *regression techniques*, will determine values for these constants c_i that best fit with the measurements. In essence, a plane will be fitted through the measurements that makes the smallest overall error with respect to the original measurements.

In Figure 4.19, we have used the same set of point measurements, but using four different approximation functions. Part (a) has indeed been determined under the assumption that the field can be approximated by a tilted plane, in this case with a downward slope from northwest to southeast. The values found by regression techniques were: $c_1 = -1.83934$, $c_2 = 1.61645$ and $c_3 = 70.8782$, giving us:

$$f(x, y) = -1.83934 \cdot x + 1.61645 \cdot y + 70.8782.$$

Clearly, not all fields are representable as simple, tilted planes. Sometimes, the theory of the application domain will dictate that the best approximation of

the field is a more complicated, higher-order polynomial function, for instance. Three classes of such functions were the basis for the fields illustrated in Figure 4.19(b)–(d).

The simplest extension from a tilted plane, that of *bilinear saddle*, expresses some dependency between the x and y dimensions:

$$f(x,y) = c_1 \cdot x + c_2 \cdot y + c_3 \cdot xy + c_4.$$

It is illustrated in part (b). A further step up the ladder of complexity is to consider *quadratic surfaces*, described by:

$$f(x,y) = c_1 \cdot x^2 + c_2 \cdot x + c_3 \cdot y^2 + c_4 \cdot y + c_5 \cdot xy + c_6.$$

The technique must now find six values for our coefficients that best match with the measurements. A bilinear saddle and a quadratic surface have been fitted through our measurements in Figure 4.19(b) and (c), respectively.

Observe that the simple, tilted plane is a special case of both a bilinear saddle and a quadratic surface, via an appropriate choice of coefficients c_i being zero. This means that if we try to approximate a field by a quadratic surface, and it is, by measurements, a perfect tilted plane, the regression techniques will just find zero values for the respective constants, thereby simplifying the formula.

Part (d) of the figure, finally, illustrates the most complex formula that we discuss here, the *cubic surface*. It is characterized by the following formula:

$$\begin{aligned} f(x,y) \;=\; & c_1 \cdot x^3 + c_2 \cdot x^2 + c_3 \cdot x + \\ & c_4 \cdot y^3 + c_5 \cdot y^2 + c_6 \cdot y + \\ & c_7 \cdot x^2 y + c_8 \cdot xy^2 + c_9 \cdot xy + c_{10}. \end{aligned}$$

The regression techniques applied for Figure 4.19 determined the following values for the coefficients c_i:

Fig 4.19	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
(a)	-1.83934	1.61645	70.8782							
(b)	-5.61587	-2.95355	0.993638	89.0418						
(c)	0.000921084	-5.02674	-1.34779	7.23557	0.813177	76.9177				
(d)	-0.473086	6.88096	31.5966	-0.233619	1.48351	-2.52571	-0.115743	-0.052568	2.16927	96.8207

Trend surface fitting is a useful technique of continuous field approximation, though determining the 'best fit' values for the coefficients c_i is a time-consuming operation, especially with many point measurements. Once these best values have been determined, we know the formula, and it has become easy to compute an approximated value for any location in the study area.

Global trends The technique of trend surface fitting discussed above can be used for the entire study area. In many cases, however, it is not very realistic to assume that the entire field is representable by some polynomial formula that is a valid approximation for *all* locations. The use of trend surface fitting for the entire area is thus at the discretion of the domain expert, who knows best whether the use of a single formula makes sense.

Another issue related to this technique is that of validity and sensitivity to spatial distribution of the measured points, and presence of outliers in the measurements. All of these can have averse effects on the resulting polynomial. This is especially true for locations that are within the study area, but outside of the area within which the measurements fall. They may be subjected to a so-called *edge effect*, meaning that the values obtained from the approximation function for edge locations may be rather nonsensical. The reader is asked to judge whether such edge effects have taken place in Figure 4.19.

Local trends In many cases, the assumption of global trend surface fitting—being that a single formula can describe the field for the *entire* study area—is an unrealistic one. Capturing all the fluctuation of a natural geographic field in a

reasonably sized study area, demands polynomials of extreme orders, and these easily become computationally intractable. Moreover, not all continuous fields are differentiable fields, and since polynomial functions are differentiable, they, again, may not be the right tools.

It is for this reason, that it can be useful to partition the study area into parts that may actually be polynomially approximated. The decision of how to partition the study area must be taken with care, and must be guided by domain expertise. For instance, if the field we want to extract from the point measurements is elevation, expert knowledge should be applied to identify the mountain ridges, as these are the places where the elevation as a function is (still continuous but) non-differentiable. A ridge line would be a good candidate to use for splitting the area. Similar 'ridges' may be present in other continuous fields, and it is the experts who should point them out.

Once we have identified the parts, we may apply the trend surface fitting techniques discussed earlier, and obtain an approximation polynomial for each part.

Even if we have taken the ridge precaution, it is probably wise to ensure that as many as possible measurements were obtained precisely from the ridges. The reason is that our local polynomials together must still form a continuous function for the whole study area. This is only the case when the two adjacent parts coincide—or at least not differ too much—in the predicted values at the ridge that forms the boundary of these parts. Occasionally, the introduction of fake, yet realistic 'measurement points' will be necessary to ensure the continuity of the global function.

Obtaining the representation of a trend surface Observe that we have discussed above the identification of an approximation *function*, either a global one

or several local ones. A function, however, is not yet a data structure in a GIS. So, how do we actually materialize the polynomial function as a raster or vector data layer?

The principles are simple. If we want to obtain a raster, we must first decide on its resolution (cell size). Then, for each cell we can determine its characteristic location (either the cell's midpoint, lower-left corner or otherwise), and apply the approximation function to that location to obtain the cell's value. Observe that this can be done in a rather simple raster calculus expression, if we know the polynomial. The measurements data are all accounted for in the trend surface function.

More elaborate cell value assignments are sometimes applied to better account for all field values occurring within the cell. One technique is to take the average of the computed values for all of the cell's corner points; again this is a straightforward raster calculus expression, though a bit longer.

If it is vector data that we want, the involved techniques are more complicated. Essentially, the aim will be to produce an isoline data layer, with a chosen 'isoline resolution'. By 'isoline resolution' we mean the list of field values for which isolines must be constructed. We do not discuss the specific techniques of how to obtain them from the approximation function but mention that triangulation techniques discussed below can play a role.

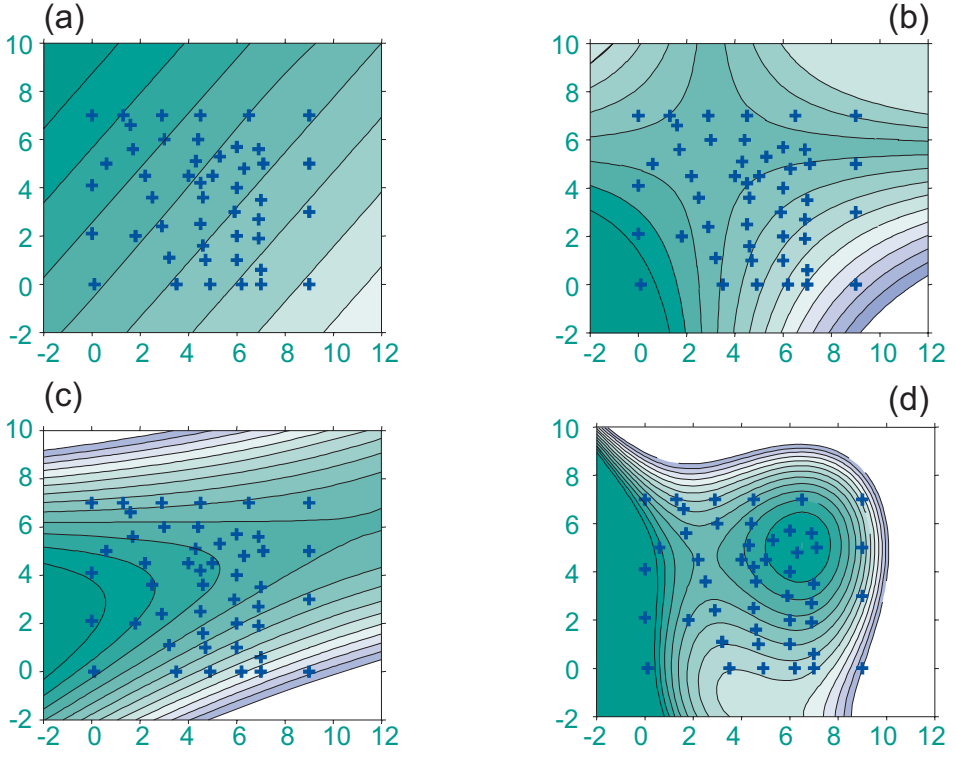


Figure 4.19: Various global trend surfaces obtained from regression techniques: (a) simple tilted plane; (b) bilinear saddle; (c) quadratic surface; (d) cubic surface.

Moving window averaging

A technique entirely different from trend surface fitting is *moving window averaging*. It too attempts to obtain a continuous field representation, this time directly into a raster data set. Moving window averaging is sometimes also called 'gridding'.

The principles behind this technique are illustrated in Figure 4.20. It computes the cell values for the output raster that represents the field one by one. To this end, a square window is defined, and initially placed over the top left raster cell. Measurement points falling inside the window contribute to the averaging computation, those outside the window do not. After the cell value is computed and assigned to the cell, the window is moved one cell to the right, and the computations are performed for that cell. Successively, all cells of the raster are visited in this way.

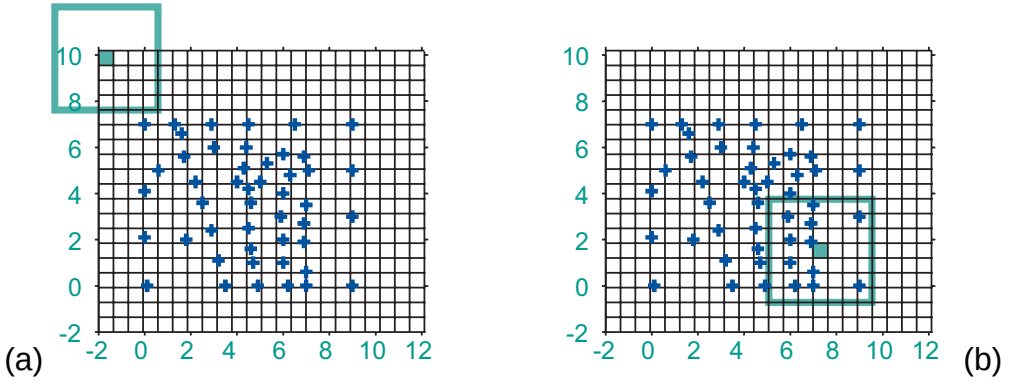


Figure 4.20: The principle of moving window averaging. In blue, the measurement points. A virtual window is moved over the raster cells one by one, and some averaging function computes a field value for the cell, using measurements within the window.

In part (b) of the figure, the 295th cell value out of the 418 in total, is being computed. This computation is based on eleven measurements, while that of the

first cell had no measurements available. Where this is the case, the cell should be assigned a value that signals this 'non-availability of measurements'.

Moving window averaging has many parameters. As a little experimentation with one's favourite GIS package will demonstrate, picking the right parameter settings may make quite a difference for the resulting raster. We discuss below the most important parameter settings.

Raster resolution Perhaps a trivial remark, but choosing an appropriate value for the raster cell size will determine whether the raster is capable of representing the field's variation. A too large cell size will smooth the function too much, removing local variations; a too small cell size will result in large clusters of equally valued cells, with little added value.

Shape/size of window Most procedures use square windows, but rectangular, circular or elliptical windows are possible too. They can be useful for instance in cases where the measurement points are distributed regularly at fixed distance over the study area, and the window shape must be chosen to ensure that each raster cell will have its window include the same number of measurement points.
The size of the window is another important matter. Small windows tend to exaggerate local extreme measurement values, for instance, statistical outliers in the measurements. Large windows have a smoothing effect on the field representation, and may negatively affect the field's variability.

Selection criteria Not necessarily all measurements within the window need to be used in averaging. Selection criteria dictate which measurements will participate in averaging and which ones will not. We may choose to use the, at most five, (nearest) measurements, or we may choose to only generate a field value if more than three measurements are in the window.

If slope or direction are important aspects of the field, the selection criteria may even be set in a way to ensure this. One technique, known as *quadrant sector control*, implements this by selecting measurements from each quadrant of the window, to ensure that somehow all directions are represented in the cell's computed value.

Averaging function A final choice is which function is applied to the selected measurements within the window. Suppose there are n measurements selected in a window, and that a measurement is denoted as m_i. The simplest averaging function will compute the standard average measurement as $\frac{1}{n}\sum_{i=1}^{n} m_i$. This function treats all measurements equally. If one feels—again, domain expertise is needed in this assessment—that measurements further away from the cell centre should have less impact than those nearby, a distance factor must be brought into the averaging function.

Functions that do this are called *inverse distance weighting functions*. Let us assume that the distance from measurement point i to the cell centre is denoted by d_i. Commonly, the weight factor applied in inverse distance weighting is the distance squared, and then the averaging formula becomes:

$$\sum_{i=1}^{n} \frac{m_i}{d_i^2} / \sum_{i=1}^{n} \frac{1}{d_i^2}.$$

In many cases in practice, one will have to experiment with parameter settings to obtain optimal results. If time series of measurements are made, with different measurement sets at different points in time, clearly one should stick

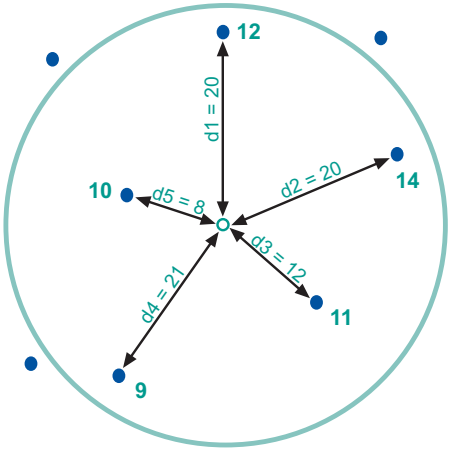


Figure 4.21: Inverse distance weighting as an averaging technique. In green, the (circular) moving window and its centre. In blue, the measurement points with their values, and distances to the centre; some are inside, some are outside of the window.

to the same parameter settings between time instants, as otherwise comparisons between fields computed for different moments in time will make little sense.

Interpolation through triangulation

Another way of interpolating point measurements is by triangulation. This technique constructs a triangulation of the study area from the known measurement points. The procedure is illustrated in Figure 4.22. Preferably, the triangulation should be a Delaunay triangulation. (For more on this type of triangulation, see Section 5.4.1.) After having obtained it, we may define for which values of the field we want to construct isolines. For instance, for elevation, we might want to have the 100 m-isoline, the 200 m-isoline, *et cetera*. For each edge of a triangle, a geometric computation can be performed that indicates which isolines intersect it, and at what positions they do. For each isoline to be constructed, this gives us a list of computed locations, all at the same field value, from which the GIS can construct the isoline. This 'spider web weaving' by the GIS is illustrated in Figure 4.22.

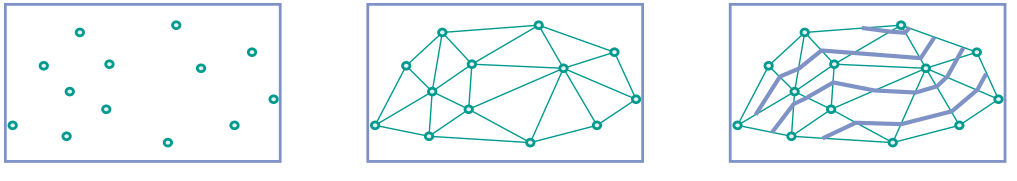


Figure 4.22: Interpolation by triangulation. (a) known point measurements; (b) constructed triangulation on known points; (c) isolines constructed from the triangulation.

4.5 Advanced operations on continuous field rasters

Continuous fields have a number of characteristics not shared by discrete fields. Since the field changes continuously, we can talk about *slope angle*, *slope aspect* and *concavity/convexity* of the slope. These notions are not applicable to discrete fields.

The discussions in this section will use terrain elevation as the prototypical example of a continuous field, but all issues discussed are equally applicable to other types of continuous fields. Nonetheless, we will regularly refer to the continuous field representation as a DEM, to conform with the commonest situation. We will assume throughout the section that the DEM is represented in a raster.

4.5.1 Applications

There are numerous examples where more advanced computations on continuous field representations are needed. We provide a short list.

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 degrees), for any or all locations.

Slope convexity/concavity calculation Slope convexity—defined as the change of the slope (negative when the slope is concave and positive when the slope is convex)—can be derived 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 in hilly and mountainous areas. The application of a special filter to a DEM produces hillshading. (For filters, see Section 4.5.2.) The colour tones in a hillshading raster represent the amount of reflected light in each location, depending on its orientation relative to the illumination source. This illumination source is usually chosen at an angle of 45° above the horizon in the north-west.

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 s/he 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) which 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 the 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 the outlet of the catchment. From there, it will repeatedly search the neighbouring pixels with the highest altitude. This process is continued until the highest location (i.e., cell with highest value) is found, and the path followed determines the catchment boundary. For delineating the drainage network, the process is reversed. Now, 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 debris flows, lava flows *et cetera*.

Visibility analysis A viewshed is the area that can be 'seen'—i.e., is in the direct line-of-sight—from a specified target location. Visibility analysis determines the area visible from a scenic lookout, the area that can be reached by a radar antenna, or assesses how effectively a road or quarry will be hidden from view.

Some of the more important of the computations mentioned above are discussed below. All of them apply a technique known as *filtering*, so we first discuss the principles of that technique.

4.5.2 Filtering

The principle of filtering is quite similar to that of *moving window averaging*, which we discussed in Section 4.4.2. Again, we define a window and let the GIS move it over the raster cell-by-cell. For each cell, the system performs some computation, and assigns the result of this computation to the cell in the output raster.

The difference with moving window averaging is that the moving window in filtering itself is a little raster, which contains cell values that are used in the computation for the output cell value. This little raster is known as the *filter*; it may be square, and commonly is, but it does not have to be. The values in the filter are often used as weight factors.

As an example, let us consider a 3×3 filter, in which all values are equal to 1, as illustrated in Figure 4.23(a). The use of this filter means that the nine cells considered are given equal weight in the computation of the filtering step. Let the input raster cell values, for the current filtering step, be denoted by r_{ij} and the corresponding filter values by w_{ij}. The output value for the cell under consideration will be computed as the sum of the weighted input values divided by the sum of weights:

$$\sum_{i,j}(w_{ij} \cdot r_{ij}) / \sum_{i,j} |w_{ij}|,$$

where one should observe we divide by the sum of *absolute* weights.

Since the w_{ij} are all equal to 1 in the case of Figure 4.23(a), the formula can be simplified to

$$\frac{1}{9}\sum_{i,j} r_{ij},$$

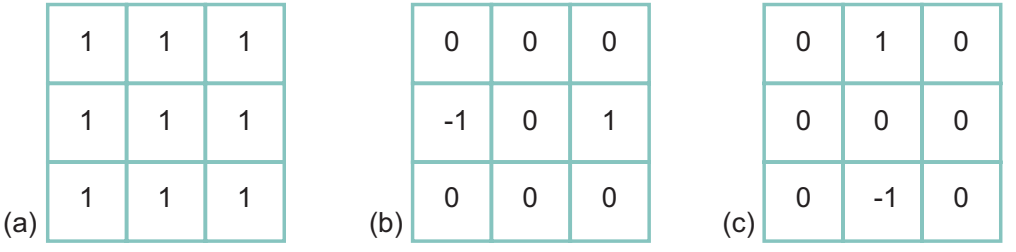


Figure 4.23: Moving window rasters for filtering. (a) raster for a regular averaging filter; (b) raster for an x-gradient filter; (c) raster for a y-gradient filter.

which is nothing but the average of the nine input raster cell values. So, we see that an 'all-1' filter computes a local average value.

More advanced filters have been devised to extract other types of information from raster data. We will look at some of these in the context of slope computations.

4.5.3 Computation of slope angle and slope aspect

Other choices 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.

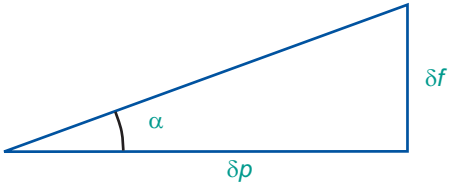


Figure 4.24: Slope angle defined. Here, δp stands for length in the horizontal plane, δf stands for the change in field value, where the field usually is terrain elevation. The slope angle is α.

Slope angle, also known as *slope gradient*, is the angle α, illustrated in Figure 4.24, made 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:

$$slope\_perc = 100 \cdot \frac{\delta f}{\delta p} \; and \; slope\_angle = \arctan(\frac{\delta f}{\delta p}).$$

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 with the North, of this maximal *down-slope* path p is what we call 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 somehow can be extracted. After all, for an arbitrary cell, we have its elevation value, plus those of its eight neighbour cells. A simple approach to slope angle computation is to make use of x-gradient and y-gradient filters.

Figure 4.23(b) and (c) illustrate an x-gradient filter, and y-gradient filter, respectively. 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 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.

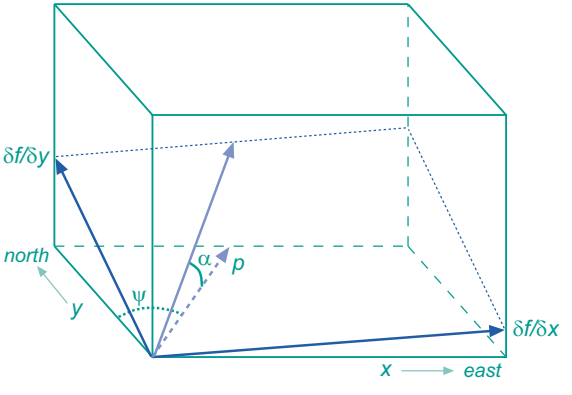


Figure 4.25: Slope angle and slope aspect defined. Here, p is the horizontal path in maximal slope direction and α is the slope angle. The plane tangent to the terrain in the origin is also indicated. The angle ψ is the slope aspect. See the text for further explanation.

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 4.25. A, not-so-simple, geometric derivation can show that always

$$\tan(\alpha) = \sqrt{(\delta f/\delta x)^2 + (\delta f/\delta y)^2}.$$

Now what does this mean in the practice of computing local slope angles from an elevation raster? It means that we must perform the following steps:

1. Compute from (input) elevation raster R the non-normalized x- and y-gradients, using the filters of Figure 4.23(b) and (c), respectively.
2. Normalize the resulting rasters by dividing by the cell width, expressed in units of length like metres.
3. Use both rasters for generating a third raster, applying the $\sqrt{\cdots}$ formula above, possibly even applying 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

$$\tan(\psi) = \frac{\delta f/\delta x}{\delta f/\delta y},$$

so slope aspect can also be computed from the normalized gradients. We must warn the reader that this formula should not trivially be replaced by using

$$\psi = \arctan(\frac{\delta f/\delta x}{\delta f/\delta y}),$$

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 $\delta f/\delta x$.)

Summary

In this chapter, we discussed some of the fundamental issues of getting spatial data into a spatial data processing system, and preparing that data for further use in analysis.

Digital data can be obtained directly from spatial data providers, or from already existing GIS application projects. A GIS project may also be involved with data obtained from ground-based surveying, which obviously have to be entered into the system. Sometimes, however, the data must be obtained from non-digital sources such as paper maps.

An issue at the heart of spatial data handling is the spatial reference system on which all the data is 'anchored'. We are too quickly believing all the time that a simple Cartesian coordinate system will do the job, but as this is a time of globalization, we should be aware about the issues related to spatial data exchange. A number of principles related to spatial reference systems, vertical and horizontal datums were discussed.

The second half of the chapter is devoted to cleaning up and further preparing data. This involves checking for errors, inconsistencies, simplification and merging existing spatial data sets. The problems that one may encounter may be caused by differences in resolution and differences in representation.

We also devoted substantial space to the issue of obtaining field representations from point measurements. This is a common problem especially in the Earth sciences, where ground-based surveys will lead to a finite list of samples that are thought to characterize a discrete or continuous geographic field. Once the field representation is obtained, one may want to perform advanced analysis like slope and aspect computations that we also discussed.

Questions

1. A colour map is scanned at the maximum resolution of a 600 dpi scanner. The map is 10×5 inches in size. A single pixel in grey-scale scanning requires one byte of storage. What will be the size of the scanned image? What will be the size if we scan it in colour mode?

2. We discussed four types of digitizing in this chapter. Which of these is the optimal one? Why? Why does (semi-)automatic digitizing require higher scanner resolutions? Under which conditions can we use it?

3. Is automatic digitizing faster than manual digitizing? Why (not)?

4. Data clean-up operations are often executed in a certain order. Why is this? Provide a sensible ordering of a number of clean-up operations.

5. Rasterization of vector data is sometimes required in data preparation. What reasons may exist for this? If it is needed, the raster resolution must be carefully selected. Argue why.

6. On page 211 in Footnote 3, we stated that 'horizontal' does not mean 'flat'. Explain this statement and refer to a figure.

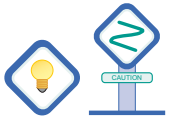
7. Assume you wish to reconcile spatial data from two neighbouring countries to resolve a border dispute. Published maps in the two countries are based on different local horizontal datums and map projections. Which steps should you take to render the data sets spatially compatible?

8. Under *Differences in accuracy* in Section 4.3.2, we looked at mapping organizations. Discuss what advantages and disadvantages exist for multi-scale databases in map production. Consider complexity of the databases involved, and consider what needs to be done if changes take place on the foundation data.

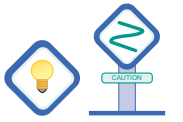
9. Take another look at Figure 4.19 and consider the determined values for the coefficients in the respective formulæ. Make a study of edge effects, for instance by computing the approximated field values for the locations $(-2, 10)$ and $(12, 10)$.

10. Figure 4.21 illustrates the technique of moving window averaging using an averaging function that applies inverse distance weighting. What field value will be computed for the cell if the averaging function is inverse *squared* distance weighting?

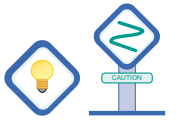
11. Construct a 3×3 window raster for filtering that approximates inverse squared distance weighting.



12. In Section 4.5, we have more or less tacitly assumed throughout to be operating on elevation rasters. All the techniques discussed, however, apply equally well to other continuous field rasters, for instance, for NDVI, population density, or groundwater salinity. Explain what slope angle and slope aspect computations mean for such fields.



13. In Section 4.5.3, we discussed *simple* x- and y-gradient filters as approximations to obtain local values for the elevation gain $\delta f/\delta x$ in x-direction, and similarly in y-direction. First explain why these filters are approximations. More advanced x- (and y-gradient) filters have been proposed in the literature. One such x-gradient filter is illustrated here.



-1	0	1
-2	0	2
-1	0	1

Figure 4.26: An advanced x-gradient filter

Explain why this filter is more advanced, and how it operates. What is the matching y-gradient filter?

Chapter 5

Spatial data analysis

Its spatio-analytic capabilities distinguish GIS from other data processing systems. These capabilities use the spatial and non-spatial data in the spatial database to answer questions and solve problems. The principal objective of spatial data analysis is to transform and combine data from diverse sources/disciplines into useful information, to improve one's understanding or to satisfy the requirements or objectives of decision-makers. A GIS application deals with only some delineated, relevant slice of reality, termed as the *universe of discourse* of the application.

Typical problems may be in planning (e.g., what are the most suitable locations for a new dam?) or in prediction (e.g., what will be the size of the lake behind the dam?). The universe of discourse here is construction of the dam, and its environmental, societal, and economic impacts. The solution to a problem always depends on a (large) number of parameters. Since these parameters are often interrelated, their interaction is made more precise in an *application*

model. Such a model, in one way or other, describes as faithfully as possible how the application's universe of discourse behaves, and it does so in terms of the parameters.[1] It is fair to say that an application model tries to simulate an application's universe of discourse.

Application models used for planning and site selection are usually *prescriptive*. They involve the use of criteria and parameters to quantify environmental, economic and social factors. The model enumerates a number of conditions to be met. In *predictive* models, a forecast is made of the likelihood of future events, which may be pollution, erosion, or even landslides. Such a model involves the expert use of various spatial data layers, either raster- or vector-based, and their combination in a methodically sound way to arrive at sensible predictions. What is 'methodically sound' to a large extent is determined by the scientific field underlying the analysis.

In this chapter, whenever we discuss spatial objects in a vector setting, we use the term 'feature' when it is immaterial whether the objects are points, lines or polygons. The topic of this chapter is analytic GIS capabilities. We first provide a classification.

[1] It is not easy to be more precise at this stage, since the nature of application models varies enormously. GIS applications for famine relief programs, for instance, are very different from earthquake risk assessment applications, though both can make use of GIS successfully.

5.1 Classification of analytic GIS capabilities

There are many ways to classify the analytic functions of a GIS. The classification used for this chapter, is essentially the one put forward by Aronoff [4]. It makes the following distinctions in function classes:

Measurement, retrieval, and classification functions allow to explore the data without making fundamental changes, and therefore they are often used at the beginning of data analysis. *Measurement functions* include computing distances between features or along their perimeters, and the computation of area size of 2D or volume size of 3D features. Counting, to understand frequency of features, is also included. *Spatial queries* retrieve features selectively, using user-defined, logical conditions. *Classification* means the (re)assignment of a thematic, characteristic value to features in a data layer.

All functions in this category are performed on single (vector or raster) data layer, often using the associated attribute data. We go in more detail in Section 5.2.

Overlay functions This group forms the core computational activity of many GIS applications. Data layers are combined and new information is derived, usually by creating features in a new layer. The computations are simpler for raster data layers than for vector layers, but both can be used. The principle of overlay is to combine features that occupy the same location.

Many GISs support overlays through an algebraic language, expressing an overlay function as a formula in which the data layers are the arguments. Different layers can be combined using arithmetic, relational, and condi-

tional operators and many different functions. Examples are provided in Section 5.3.

Neighbourhood functions Whereas overlays combine features at the same location, neighbourhood functions evaluate the characteristics of an area *surrounding* a feature's location. This allows to look at buffer zones around features, and spreading effects if features are a source of something that spreads—e.g., water springs, volcanic eruptions, sources of pollution. We discuss these topics more fully in Section 5.4.

Connectivity functions evaluate how features are connected. This is useful in applications dealing with networks of connected features. Examples are road networks, water courses in coastal zones, and communication lines in mobile telephony. Details are discussed in Section 5.5.

5.2 Retrieval, classification and measurement

5.2.1 Measurement

Geometric measurement on spatial features includes counting, distance and area size computations. For the sake of simplicity, this section 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, as discussed in Section 3.3.4.

Measurements on vector data are more advanced, thus, also more complex, than those on raster data. We discuss each group.

Measurements on vector data

The primitives of vector data sets are 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.

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; some GISs store these also, others compute them 'on-the-fly'.

Length is a geometric property associated with polylines, by themselves, or in their function as polygon boundary. It can obviously be computed by the GIS—as the sum of lengths of the constituent line segments—but it quite often is also stored with the polyline.

Area size is associated with polygon features. Again, it can be computed, but usually is stored with the polygon as an extra attribute value. This speeds up the computation of other functions that require area size values. We see that all of the above measurements do not require computation, but only a look up in stored data.

Measuring 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 are 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 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.

Observe that we cannot possibly store all distance values for all possible combinations of two features in any reasonably sized spatial database. So, the system must compute 'on the fly' whenever a distance computation request is made.

Another geometric measurement used by the GIS 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 5.1. Bounding box computation is an important support function for the GIS: 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 an expensive function, but bounding box computation is not, the GIS will always first apply the latter as a test to see whether it must do the first.

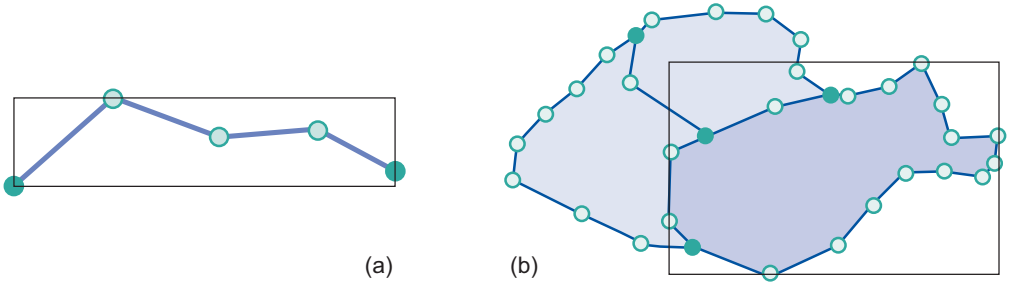


Figure 5.1: The minimal bounding box of (a) a polyline, and (b) a polygon

For practical purposes, it is important to understand what is the measurement unit in use for the spatial data layer that one operates 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 in GIS use, and we want measurements to be performed on these interactively defined features. Then, the GIS will have to perform possibly 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 do not. Together with the location of a so-called 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.

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 case of low resolution data they become more important to be aware of.

The *area size* of a selected part of the raster (a group of cells) is calculated as the number of cells multiplied with the cell area size.

The *distance* between two raster cells is the standard distance function applied to the locations of their respective mid-points, obviously taking into account the cell resolution. 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. This computation is prone to error, as we already discovered in Question 2.13.

5.2.2 Spatial selection queries

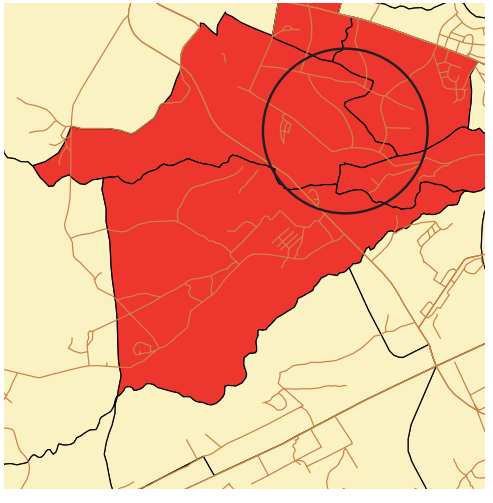
When exploring a spatial data set, the first thing one usually wants is to select certain features, to (temporarily) restrict the exploration. Such selections can be made on geometric/spatial grounds, or on the basis of attribute data associated with the spatial features. We discuss both techniques below.

Interactive spatial 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 2.14) with the selection objects. These become the *selected objects*.

As we have seen in Section 3.3.6, spatial data is usually associated with its attribute data (stored in tables) through a key/foreign key link. Selections of features lead, via these links, to selections on the records. *Vice versa*, selection of records may lead to selection of features.

Interactive spatial selection answers questions like "What is at ...?" In Figure 5.2, the selection object is a circle and the selected objects are the red polygons; they overlap with the selection object.

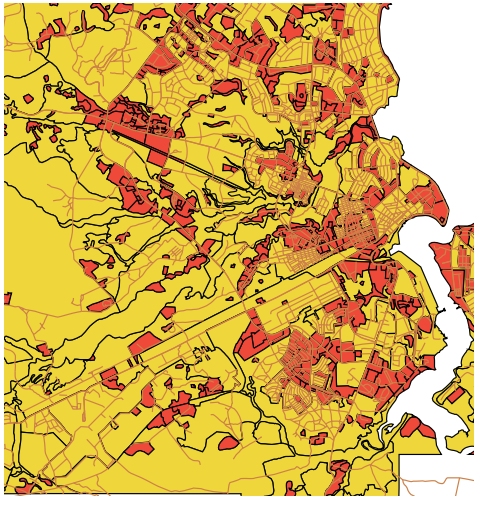


Area	Perimeter	Ward_id	Ward_nam	District	Pop88	Pop92
65420380.0000	41654.940000	1	KUNDUCHI	Kinondoni	22106	27212.00
24813620.0000	30755.620000	2	KAWE	Kinondoni	32854	40443.00
18698500.0000	26403.580000	3	MSASANI	Kinondoni	51225	63058.00
81845610.0000	49645.160000	4	UBUNGO	Kinondoni	47281	58203.00
4468546.00000	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

Figure 5.2: 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 on Dar es Salaam, Tanzania. Data source: Division of Urban Planning and Management, ITC.

Spatial selection by attribute conditions

One can also select features by stating selection conditions on the features' 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 . . . ?"



Area	IDs	LandUse
174308.7000	2	30
2066475.000	3	70
214582.5000	4	80
29313.8600	5	80
73328.0800	6	80
53303.3000	7	80
614530.1000	8	20
1637161.000	9	80
156357.4000	10	70
59202.2000	11	20
83289.5900	12	80
225642.2000	13	20
28377.3300	14	40
228930.3000	15	30
986242.3000	16	70

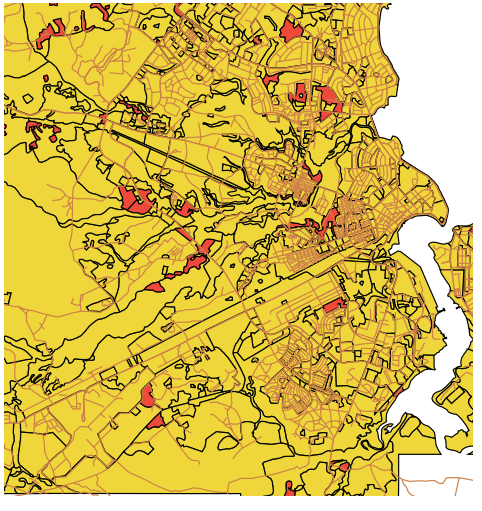
Figure 5.3: Spatial selection using the attribute condition $Area < 400000$ on land use areas in Dar es Salaam. Spatial features on left, associated attribute data (in part) on right. Data source: Division of Urban Planning and Management, ITC.

Figure 5.3 shows an example of selection by attribute condition. The query expression is $Area < 400000$, which can be interpreted as "select all the land use areas of which the size is less than $400,000$." The polygons in red are the selected areas; their associated records are also highlighted in red.

We can use an already selected set of features as the basis of 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 5.3 are subjected to a further condition, $LandUse = 80$. The result is illustrated in Figure 5.4.

Such combinations of conditions are fairly common in practice, so we devote a small paragraph on the theory of combining conditions.



Area	IDs	LandUse
174308.7000	2	30
2066475.000	3	70
214582.5000	4	80
29313.8600	5	80
73328.0800	6	80
53303.3000	7	80
614530.1000	8	20
1637161.000	9	80
156357.4000	10	70
59202.2000	11	20
83289.5900	12	80
225642.2000	13	20
28377.3300	14	40
228930.3000	15	30
986242.3000	16	70

Figure 5.4: Further spatial selection from the already selected features of Figure 5.3 using the additional condition $LandUse = 80$ on land use areas. Observe that fewer features are now selected. Data source: Division of Urban Planning and Management, ITC.

Combining attribute conditions

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*.

Above, we have seen simple, atomic conditions such as $Area < 400000$ and $LandUse = 80$. 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, to form an atomic condition. For instance, $LandUse <> 80$ can be used to select all areas with a land use class different from 80. Expressions are either constants like 400000 and 80, attribute names like $Area$ and $LandUse$, or possibly composite arithmetic expressions like $0.15 \times Area$, which would compute 15% of the area size.

Atomic conditions can be combined into composite conditions using *logical connectives*. The most important ones to know—and the only ones we discuss here—are AND, OR, NOT and the bracket pair $(\cdots)$. If we write a composite condition like

$$Area < 400000 \; AND \; LandUse = 80,$$

we are selecting areas for which *both* atomic conditions hold. This is the semantics of the AND connective. If we had written

$$Area < 400000 \; OR \; LandUse = 80$$

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 $NOT\ (LandUse = 80)$ would select all areas with a different land use class than 80. (Clearly, the same selection can be obtained by writing $LandUse <> 80$, 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

$$(Area < 30000\ AND\ LandUse = 70)\ OR\ (Area < 400000\ AND\ LandUse = 80)$$

will select areas of class 70 less than $30,000$ in size, as well as class 80 areas less than $400,000$ in size.

Spatial selection using topological relationships

Various forms of topological relationship between spatial objects were discussed in Section 2.2.4. These relationships can be useful to select features as well. We will look at containment, overlap, neighbourhood and also at selections on the basis of a distance function. The steps carried out are always

1. to select one or more features as the selection objects, and
2. to apply the 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 5.5 illustrates a containment query. Here, we were interested in finding out where are the medical clinics in the area of Ilala District. We first selected all areas of Ilala District, using the technique of selection by attribute condition $District = \text{“}Ilala\text{”}$. Then, these selected areas were used as selection objects to determine which medical clinics (as point objects) were within them.

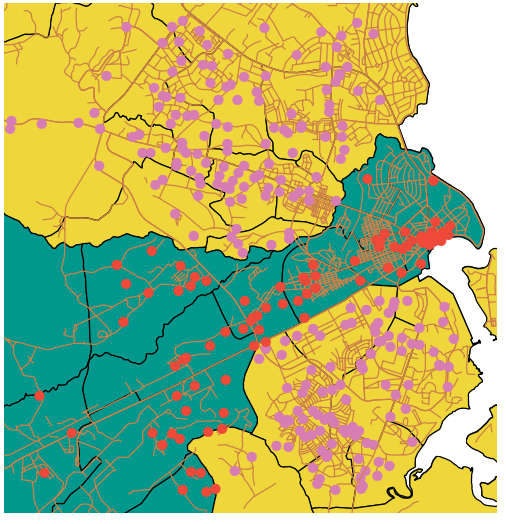


Figure 5.5: Spatial selection using containment. In dark green, all wards within Ilala District as the selection objects. In red, all medical clinics located inside these areas, and thus inside the district. Data source: Division of Urban Planning and Management, ITC.

Selecting features that intersect The intersect operator identifies features that are not disjoint in the sense of Figure 2.14, but extended to points and lines. Figure 5.6 provides an example of spatial selection using the intersect relationship between lines and polygons. We selected all roads intersecting Ilala District.

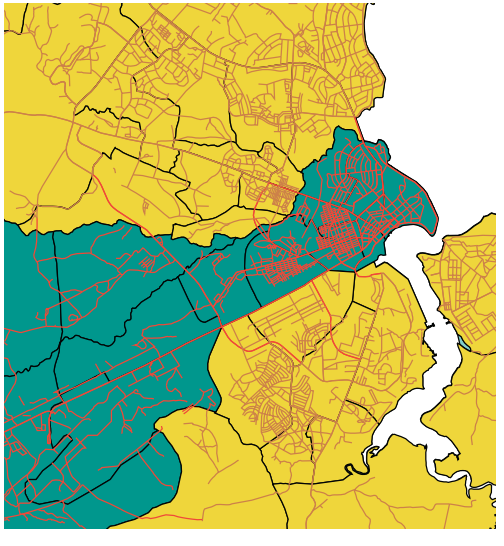


Figure 5.6: Spatial selection using intersection. The wards of Ilala District function as the selection objects (in dark green), and all roads (partially) in the district are selected (in red). Data source: Division of Urban Planning and Management, ITC.

Selecting features adjacent to selection objects Adjacency is the meet relationship of Section 2.2.4. It expresses that features share boundaries, and therefore it applies only to line and polygon features.

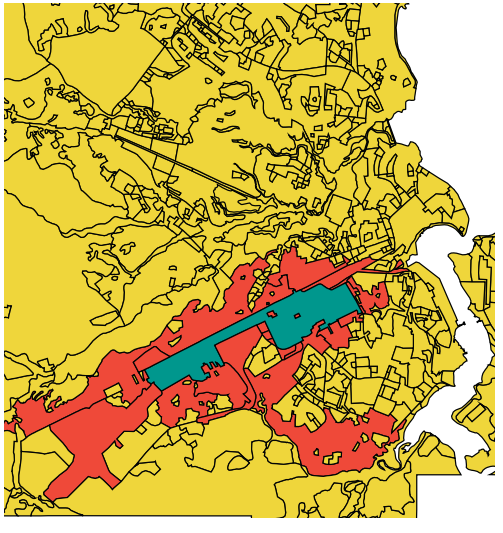


Figure 5.7: Spatial selection using adjacency. Our selection object is an industrial area near down town Dar es Salaam, Tanzania; our adjacency selection finds all adjacent land use areas. Data source: Division of Urban Planning and Management, ITC.

Figure 5.7 illustrates a spatial adjacency query. We want to select all parcels adjacent to an industrial area. The first step is to select that area (in dark green) and then apply the adjacency function to select all land use areas (in red) that are adjacent to it.

Selecting features based on their distance One may also want to use the distance function of the GIS as a tool in selecting features. Such selections can be searches *within* a given distance from the selection objects, *at* a given distance, or

even *beyond* a given distance. There is a whole range of applications to this type of selection:

- Which clinics are within 2 kilometres of a selected school? (Information needed for the school emergency plan.)
- Which roads are within 200 metres of a medical clinic? (These roads must have a high road maintenance priority.)

Figure 5.8 illustrates a spatial selection using distance. Here, we executed the selection of the second example above. Our selection objects were all clinics, and we selected the roads that pass by a clinic within 200 metres.

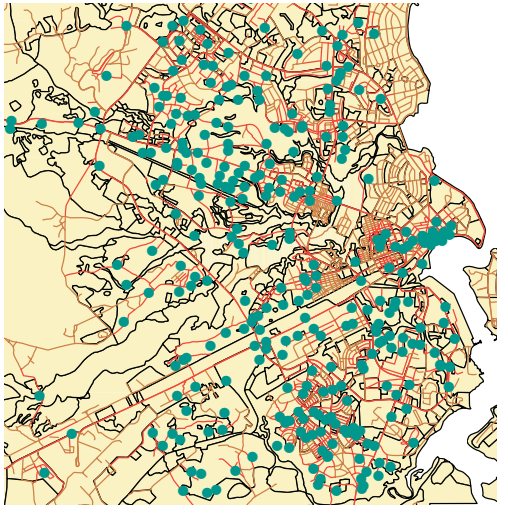


Figure 5.8: Spatial selection using the distance function. With all clinics being our selection objects, we searched for roads that pass by within 200 metres. Observe that this also selects road segments that are far away from any clinic, simply because they belong to a road of which a segment is nearby. Data source: Division of Urban Planning and Management, ITC.

In situations in which we know what distance value to use—for selections within, at or beyond that distance value—the GIS has many (straightforward)

computations to perform. Things become more complicated if our 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 it must find that feature F that is nearest to (resp., 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.

Afterthought on selecting features We have now discussed a number of different techniques for selecting features. We have also seen that selection conditions on attribute values can be combined using logic connectives like AND, OR and NOT. A fact is that the other techniques of selecting features are usually combinable *as well*. 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 the roads within 200 metres, then select from them only the major roads, then select the nearest clinics to these remaining roads, as the ones that should receive our financial support. Essentially, we are combining in this way various techniques of selection.

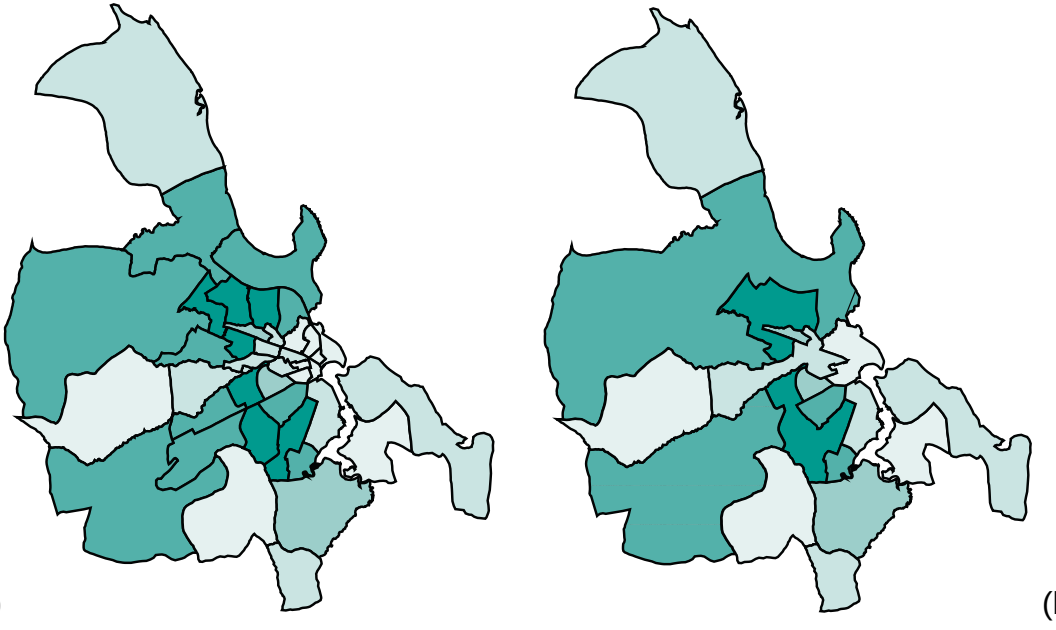
5.2.3 Classification

Classification is a technique of purposefully 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. We do so by assigning a characteristic value to each element in the input set—which is usually a collection of spatial features that can be raster cells or points, lines or polygons. If the number of characteristic values is small in comparison to the size of the input set, we have *classified* the input set.

The pattern that we look for may be the distribution of household income in a city. 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 in a sensible colour scheme, this may reveal interesting information. This has been done for Dar es Salaam in Figure 5.9 in two ways.

The input data set may have been itself the result of some classification, and in such a case we talk of 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.

In classification of vector data, there are two possible results. 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 spatial ex-



(a) (b)

Figure 5.9: Two classifications of average annual household income per ward in Dar es Salaam, Tanzania. Higher income areas in darker greens. Five categories were identified. (a) with original polygons left intact; (b) with original polygons merged when in same category. The data used for this illustration are *not* factual.

tents of the original features. Figure 5.9(a) is an illustration of this first type of output. A second type of output is obtained when adjacent features with 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 5.9(b). 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 is sensible only for lines and polygons.

Below, we discuss two kinds of classification: user-controlled and automatic.

User-controlled classification

In *user-controlled classification*, we indicate which attribute is, or which ones are, the classification parameter(s) and we define the classification method. The latter involves declaring the number of classes as well as the correspondence between the old attribute values and the new classes. This is usually done via a classification table. The classification table used for Figure 5.9 is displayed in Table 5.1. It is rather typical for cases in which the used parameter domain is continuous (as in household income). Then, the table indicates *value ranges* to be mapped to the same category. Observe that categorical values are ordinal data, in the sense of Section 2.1.3.

Household income range	*New category value*
391–2474	1
2475–6030	2
6031–8164	3
8165–11587	4
11588–21036	5

Table 5.1: Classification table used in Figure 5.9.

Another case exists when the classification parameter is nominal or at least discrete. Such an example is given in Figure 5.10.

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 is the data type of the attribute(s) from which the classification was obtained.

Sometimes, one may want to perform classification only on 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

Code	Old category	New category
10	Planned residential	Residential
20	Industry	Commercial
30	Commercial	Commercial
40	Institutional	Public
50	Transport	Public
60	Recreational	Public
70	Non built-up	Non built-up
80	Unplanned residential	Residential

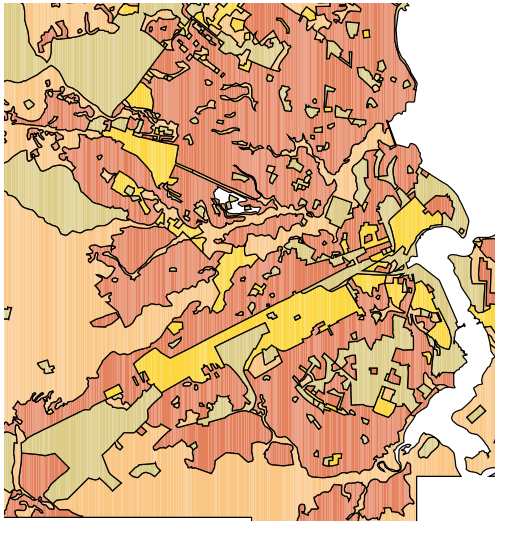


Figure 5.10: An example of a classification on a discrete parameter, namely land use unit in the city of Dar es Salaam, Tanzania. Colour scheme: Residential (brown), Commercial (yellow), Public (Olive), Non built-up (orange). Data source: Division of Urban Planning and Management, ITC.

value to them in the output data set. A null value is a special value that means that no applicable value is present. Care must be taken to deal with these values correctly, both in computation 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. Two techniques of determining break points are in use.

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 revealing the distribution patterns 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 and by 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 5.11.

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

(a) original raster

1	1	1	1	4
2	2	3	2	5
2	2	2	1	5
2	3	3	4	4
2	1	1	1	1

(b) equal interval 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

1	1	1	2	5
3	3	4	3	5
3	2	2	2	5
3	4	4	5	5
3	2	1	1	1

(c) equal frequency classification

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 5.11: 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 the number of cells involved.

5.3 Overlay functions

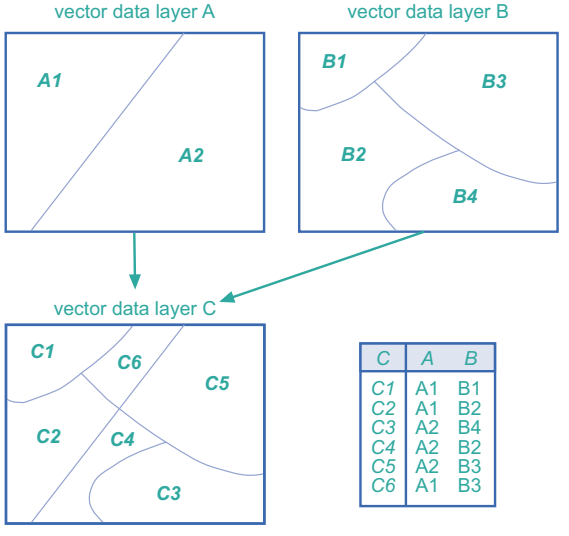
In the previous section, we saw various techniques of measuring and selecting spatial data. We also discussed the generation of a new spatial data layer from an old one, using classification. In this section, we look at techniques of combining two spatial data layers and producing a third one from them. The binary operators that we discuss are known as *spatial overlay operators*. We will first discuss vector forms, and then raster overlay operators.

Standard overlay operators take two input data layers, and assume they are georeferenced in the same system, and overlap in study area. If either condition is not met, the use of an overlay operator is senseless. The principle of spatial overlay is to compare the characteristics of the same location in both data layers, and to produce a new characteristic for each location in the output data layer. Which characteristic to produce is determined by a rule that the user can choose.

In raster data, as we shall see, these comparisons are carried out between pairs of cells, one from each input raster. In vector data, the same principle of comparing locations pairwise applies, but the underlying computations rely on determining the spatial intersections of features, one from each input vector layer, pairwise.

5.3.1 Vector overlay operators

In the vector domain, the overlaying of data layers is computationally more demanding than in the raster domain. We will discuss here only overlays from polygon data layers, but remark that most of the ideas carry over to overlaying with point or line data layers.



C	A	B
C1	A1	B1
C2	A1	B2
C3	A2	B4
C4	A2	B2
C5	A2	B3
C6	A1	B3

Figure 5.12: The polygon intersect overlay operator. Two polygon layers A and B produce a new polygon layer (with associated attribute table) that contains all intersections of polygons from A and B. Figure after [9].

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 5.12. 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 Chapter 3—of the two input attribute tables. This

output attribute table only contains a tuple for each intersection polygon found, and this explains why we call this operator sometimes a *spatial join*.

A more practical example is provided in Figure 5.13, which was produced by polygon intersection of the ward polygons with land use polygons classified as in Figure 5.10. This has allowed us to select the residential areas in Ilala District.

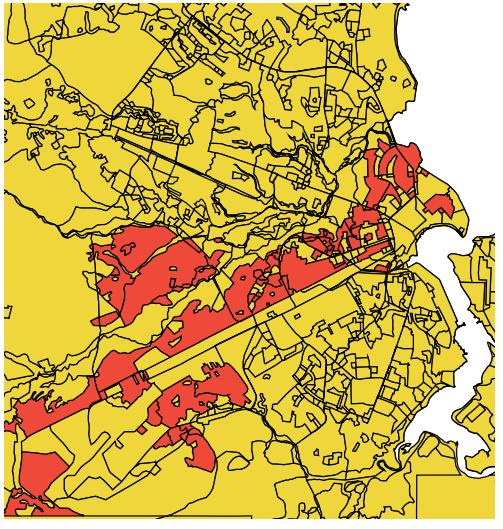


Figure 5.13: The residential areas of Ilala District, obtained from polygon intersection. Input for the polygon intersection operator were (a) a polygon layer with all Ilala wards, (b) a polygon layer with the residential areas, as classified in Figure 5.10. Data source: Division of Urban Planning and Management, ITC.

Two more polygon overlay operators are illustrated in Figure 5.14. The first 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 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.

A second overlay operator is *polygon overwrite*. The result of this binary operator is defined as a polygon layer with the polygons of the first layer, except

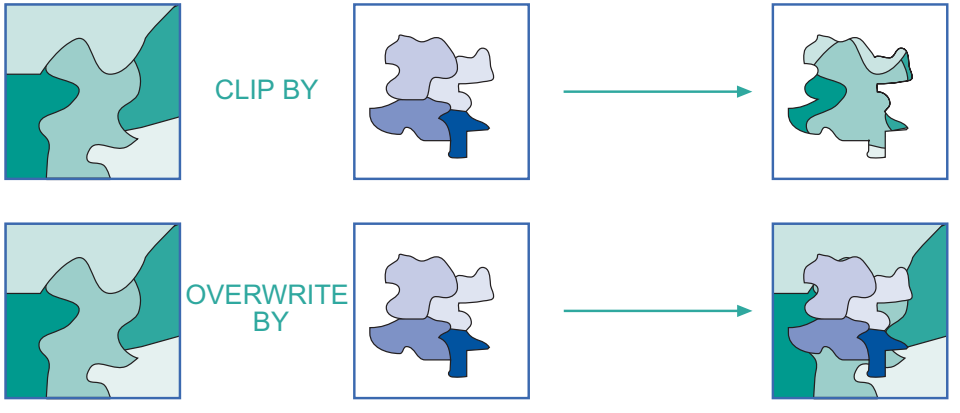


Figure 5.14: Two more polygon overlay operators: (a) polygon clip overlay clips down the left hand polygon layer to the generalized spatial extent of the right hand polygon layer; (b) polygon overwrite overlay overwrites the left hand polygon layer with the polygons of the right hand layer.

where polygons existed in the second layer, as these take priority. The principle is illustrated in the lower half of Figure 5.14.

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 of polygons as operator argument.

The really fundamental operator of all these is polygon intersection. The others 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 also defined usually for point or line data layers. Their definition parallels the definitions of operators discussed above. Different GISs use different names for these operators, and one is advised to carefully check the

documentation before applying any of these operators.

5.3.2 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 do most—usually have a full language to express operations on rasters. We could call such a language a *raster calculus*, as it allows to compute new rasters from existing ones, using a range of functions and operators. Unfortunately, raster calculi come disguised under various names, and not all offer the same functionality. ILWIS's raster calculus is known as the map algebra, for instance. The discussion below is to a large extent based on ILWIS's raster calculus, though in general terminology.

When producing a new raster we must provide a name for it, and define how it is computed. This is done in an assignment statement of the following format:

$$Output\_raster\_name := Raster\_calculus\_expression.$$

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 below. 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 until the last pixel in the last row.

There is a wide range of operators and functions that can be used in raster calculus.

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 $10 - 9 = 1$. Similarly, $10\ DIV\ 3$ will return 3. More operators are goniometric: sine (sin), cosine (cos), tangent (tan), and their inverse functions $asin$, $acos$, and $atan$, which return radian angles as real values.

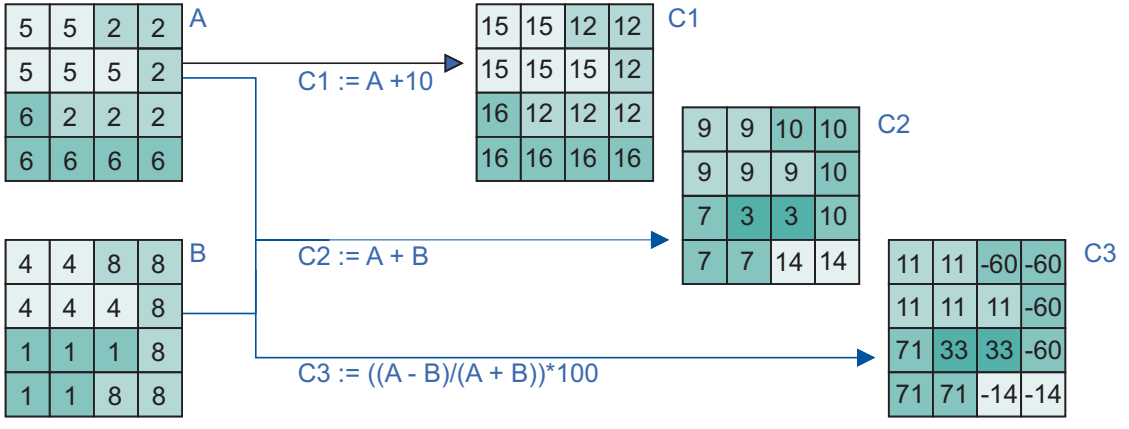


Figure 5.15: Examples of arithmetic raster calculus expressions

Some simple raster calculus assignments are illustrated in Figure 5.15. 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.

Comparison and logical operators

Raster calculus also allows to compare rasters, cell by cell. To this end, we may use the standard comparison operators ($<$, $<=$, $=$, $>=$, $>$ and $<>$) that we introduced before.

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 Section 5.2.2. Another connective that is commonly offered in raster calculus is *exclusive OR* (XOR). The expression $a\ XOR\ b$ is true if either a or b is true, but not both.

Examples of the use of these comparison operators and connectives are provided in Figure 5.16 and Figure 5.17. The latter figure provides various raster computations in search of forests at specific elevations.

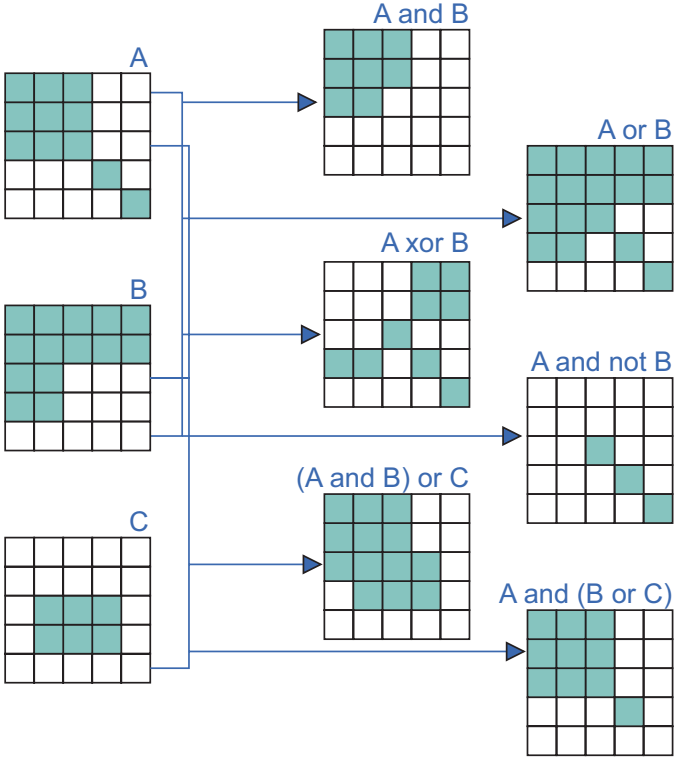


Figure 5.16: Examples of logical expressions in raster calculus. Green cells represent true values, white cells represent false values.

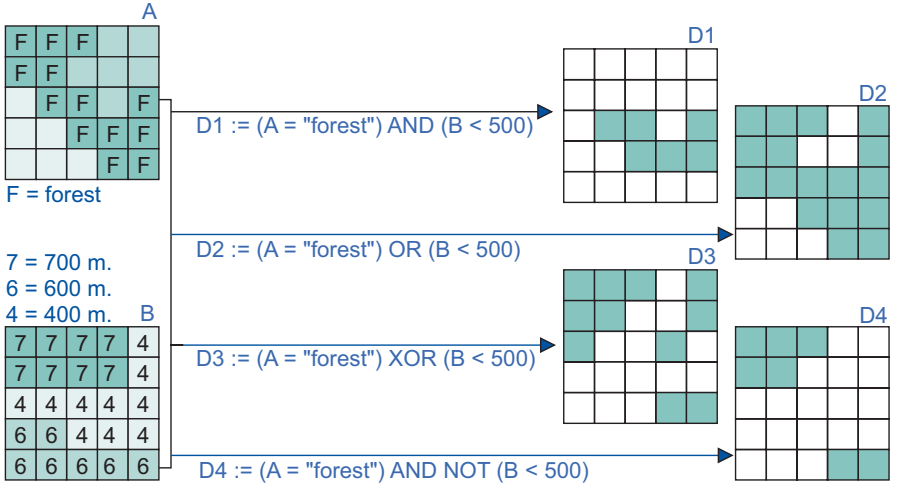


Figure 5.17: Examples of complex logical expressions in raster calculus. Raster A is a classified raster for land use, raster B one with elevation values. Raster $D1$ indicates where is forest below 500 m, raster $D2$ indicates areas below 500 m and forests, raster $D3$ areas that are either forest or 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 with them that allows to test whether a condition is fulfilled. The general format is:

$$Output\_raster := IFF(condition, then\_expression, else\_expression).$$

Here, *condition* is the tested condition, *then_expression* is evaluated if *condition* holds, and *else_expression* is evaluated if it does not hold. This means that an expression like $IFF(4 = 5, \text{“}land\text{”}, \text{“}lake\text{”})$ will evaluate to “*lake*” since $4 = 5$ is not true, so the *else_expression* is evaluated. More examples of are provided in Figure 5.18.

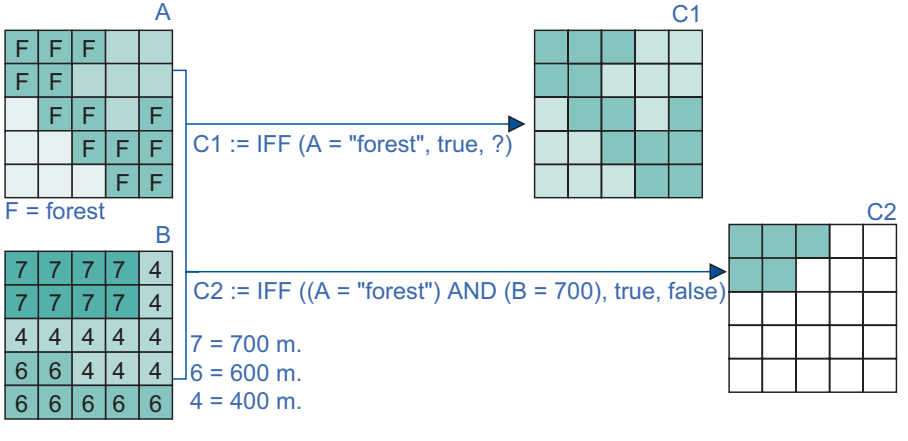


Figure 5.18: Examples of conditional expressions in raster calculus. A is a classified raster holding land use data, and B is an elevation value raster. The expression indicated as ‘?’ represents an ‘unknown’ truth value, and is depicted in light green.

5.3.3 Overlays using a decision table

Conditional expressions are powerful tools in cases where multiple criteria must be taken into account. A small size 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 illustrated in Figure 5.19 on the left. 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 the others are not.

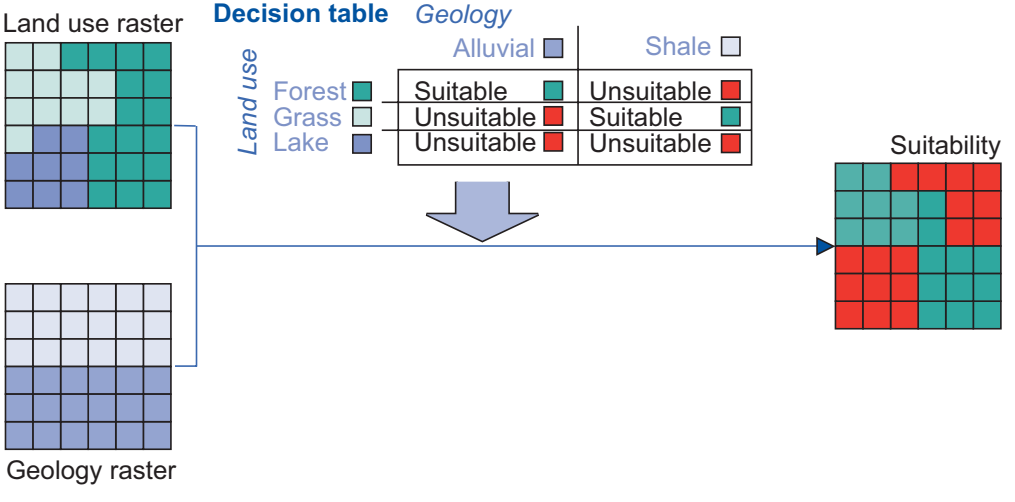


Figure 5.19: The use of a decision table in raster overlay. The overlay is computed in a suitability study, in which land use and geology are important factors. The meaning of values in both input rasters, as well as the output raster can be understood from the decision table.

We could produce the output raster of Figure 5.19 with a longish raster calculus expression like

$$\text{Suitability} := IFF((Landuse = \text{“}Forest\text{”}\ AND\ Geology = \text{“}Alluvial\text{”})\ OR$$

$$(Landuse = \text{“}Grass\text{”}\ AND\ Geology = \text{“}Shale\text{”}),$$
$$\text{“Suitable”, “Unsuitable”})$$

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 *IFF* 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 5.19. The GIS will have supporting functions to generate the additional table from the input rasters, and to enter appropriate values in the table.

5.4 Neighbourhood functions

In our section on overlay operators, 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 raster calculus, for instance, gave 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 here 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* the 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 what is their spatial extent,
2. define how to determine the neighbourhood for each target,
3. define which characteristic(s) must be computed for each neighbourhood.

For instance, our target can be a medical clinic. Its neighbourhood can be defined as

- an area within 2 km distance, as the crow flies, or
- an area within 2 km travel distance, or
- all roads within 500 m travel distance, or
- all other clinics within 10 minutes travel time, or

- all residential areas, for which the clinic is the closest clinic.

Then, in the third step we indicate what characteristics to find out about the neighbourhood. This could simply be its spatial extent, but it might also be statistical information like

- how many people live in the area,
- what is their average household income, or
- are any high-risk industries located in the neighbourhood.

The above 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 then, neighbourhood characteristics often are obtained via statistical summary functions that compute values such as 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 5.2.2. To obtain characteristics from an eventually 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. We discuss some of these techniques in Section 5.4.1. Geometric distance does not take into account direction and certain phenomena can only be studied by doing so. Think of pollution spread by rivers, ground water flow, or prevailing weather systems. The more advanced techniques for these are covered as spread computations in Section 5.4.2. Spread functions are based on the assumption that the phenomenon spreads in *all* directions, though not necessarily equally easily in all directions. Hence, it uses

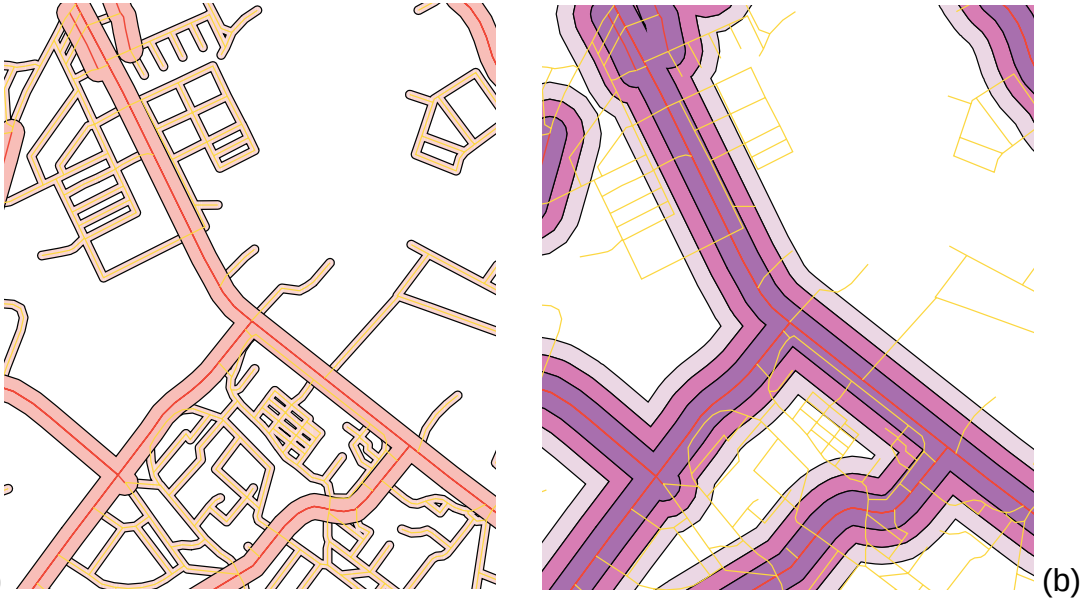
local terrain characteristics to compute the local resistance against spreading. In seek computations, the assumption is that the phenomenon will choose a least-resistance path, and *not* spread in all directions. This, as we will see, involves the computation of preferred local direction of spread. Spread and seek computations take local characteristics into account, and thus are better performed in raster data.

5.4.1 Proximity computation

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 we discuss 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 5.20(a), a number of main and minor roads were selected as targets, and a 75 m (resp., 25 m) buffer was computed from them.



(a) (b)

Figure 5.20: Buffer zone generation: (a) around main and minor roads. Different distances were applied: 25 metres for minor roads, 75 metres for main roads. (b) Zonated buffer zones around main roads. Three different zones were obtained: at 100 metres from main road, at 200, and at 300 metres.

In some case studies, zonated buffers must be determined, for instance in assessments of traffic noise effects. Most GIS support this type of zonated buffer computations. An illustration is provided in Figure 5.20(b).

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.

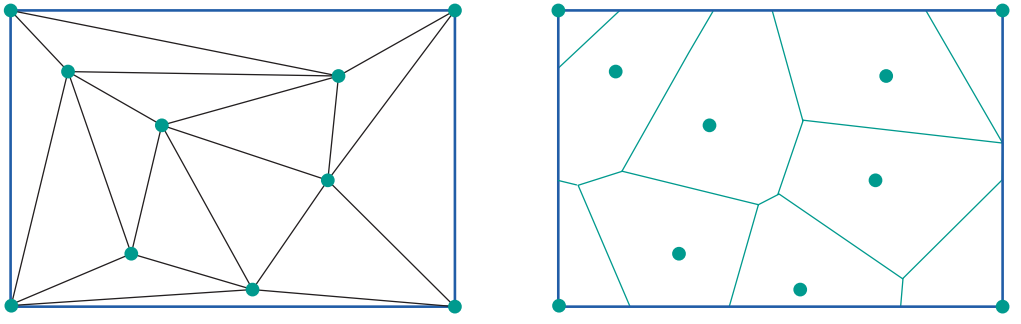


Figure 5.21: Thiessen polygon construction from a Delaunay triangulation: perpendiculars of the triangles form the boundaries of the polygons.

Thiessen polygon generation

Another technique that makes use of geometric distance for determining neighbourhoods is useful if we have a spatially distributed set of points as target locations, and we want to know for each location in the study to which target it is closest. This technique will generate a polygon around each target location that identifies all those locations that 'belong to' that target.

A partitioning of the plane into polygons that have this characteristic—containing all the locations that are closer to the polygon's 'midpoint' than to any other 'midpoint'—is called a Thiessen polygon partition. 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 Section 2.2.3 on TINs).

Figure 4.18(a) repeats the Delaunay triangulation of Figure 2.8(b). The Thiessen polygon partition constructed from it is in part (b). The construction first creates the perpendiculars of all the triangle sides; observe that a perpendicular of a triangle side that connect point A with point B is the divide between the area closer to A and the area closer to B. The perpendiculars become part of the

boundary of each Thiessen polygon.

5.4.2 Spread computation

The determination of neighbourhood of one or more target locations may depend not only on distance—cases which we discussed above—but also on direction and differences in the terrain in different directions. This typically is the case when the target location contains a 'source material' that spreads over time. This 'source material' may be air, water or soil pollution, commuters exiting a train station, people from an opened-up refugee camp, a water spring uphill, or the 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 terrain factors that influence the spread, making it easier or more difficult. Many GIS provide support for this type of *spread computation*, and we discuss some of its principles, in the context of raster data, here.

Spread computation involves one or more target locations, which are better called *source locations* in this context. They are the locations of the source of whatever spreads. Spread 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 by 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 witnessed for reaching a raster cell. All of this is illustrated in Figure 5.22.

While computing total resistances, the GIS takes proper care of correct spread path lengths. Obviously, the spread from a cell c_{src} to its neighbour cell to the east c_{e} is shorter than to the cell that is its northeast 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 spreading from c_{src} to

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

(a)

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

(b)

Figure 5.22: Spread computations 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 higher precision real arithmetic than what is illustrated here.)

c_{e} as $\frac{1}{2}(val(c_{\mathrm{src}}) + val(c_{\mathrm{e}}))$, while the same for c_{src} to c_{ne} is $\frac{1}{2}(val(c_{\mathrm{src}}) + val(c_{\mathrm{ne}})) \times \sqrt{2}$. 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 5.22(b).

For instance, there are three paths from the green source location to its north-east 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:

$$\frac{1}{2}(4+4)+\frac{1}{2}(4+5)=8.5.$$

Path 2, in similar style, gives us a total value of 6.5. For path 3, we find

$$\frac{1}{2}(4+5)\times\sqrt{2}=6.36,$$

and thus it obviously is the minimal cost path. The reader is asked to verify one or two other values of minimal cost paths that the GIS has produced.

5.4.3 Seek computation

Spread computations determine how a phenomenon spreads over the area, in principle in all directions, though with different difficulty or resistance. There are, however, also cases where a phenomenon does not spread in all directions, but only along a chosen, least-cost path, determined again by local terrain characteristics. The typical case arises when we want to determine the drainage patterns in a catchment: the rainfall water 'chooses' a way to leave the area. This is when we use *seek computations*.

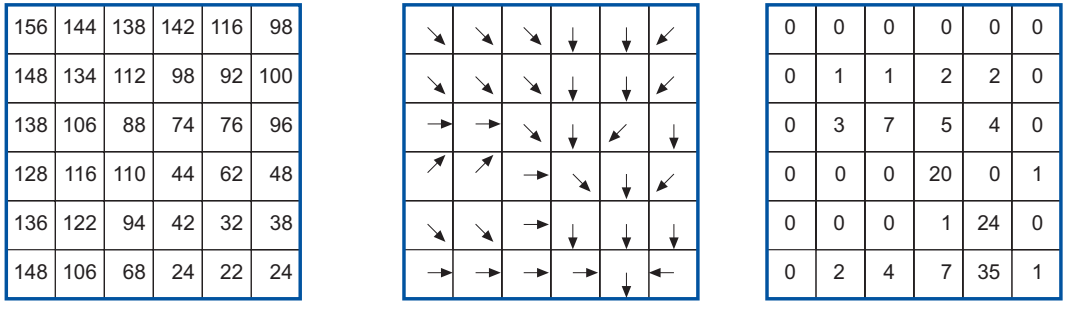


Figure 5.23: Seek computations on a raster: (a) the original elevation raster, (b) the flow direction raster computed from it, (c) accumulated flow count raster.

We illustrate the principles with a simple elevation raster, provided in Figure 5.23(a). For each cell in that raster, the steepest downward slope to a neighbour cell is computed, and its direction stored in a new raster (Figure 5.23(b)). This computation determines the elevation difference between the cell and a neighbour cell, and takes into account cell distance—1 for neighbour cells in N–S or W–E direction, $\sqrt{2}$ for cells in NE–SW or NW-SE direction. Among its eight neighbour cells, it picks the one with the steepest path to it. The directions in raster (b), thus obtained, are encoded in integer values, and 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 the cell.

Cells with a high accumulated flow count represent areas of concentrated flow, and thus may belong to a stream. By using some appropriately chosen threshold value in a raster calculus expression, we may decide whether they do. Cells with an accumulated flow count of zero are local topographic highs, and can be used to identify ridges.

5.5 Network analysis

A completely different set of analytic functions in GIS consists of computations on networks. A *network* is a connected set of lines, representing some geographic phenomenon, typically of the transportation type. 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 done using either raster or vector data layers, but they are more commonly done in the latter, as line features can be associated with a network naturally, and can be given typical transportation characteristics like capacity and cost per unit. One crucial characteristic of any network is whether the network lines are considered 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.)

For many applications of network analysis, a *planar network*, i.e., one that is embeddable in a two-dimensional plane, will do the job. Many networks are naturally planar, like stream/river networks. A large-scale traffic network, on the other end, 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 can do so only using trickery.

Such trickery may involve to split overpassing lines at the intersection vertex and create four out of the two original lines. Without further attention, the net-

work will then allow to make a turn onto another line at this new intersection node, which in reality would be impossible. Some GIS allow to associate a cost with 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 we said, this is trickery 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 GIS provide rule-based tools that allow the definition of these extra application rules.

Various classical spatial analysis functions on networks are supported by GIS software packages. The most important ones are:

Optimal path finding which generates a least cost-path on a network between a pair of predefined locations using both geometric and attribute data.

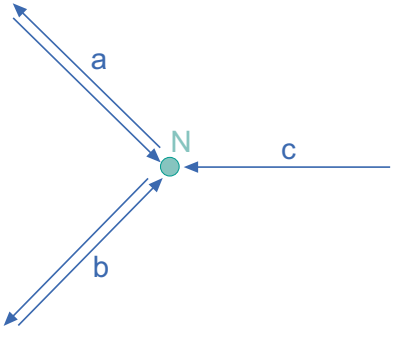
Network partitioning which assigns network elements (nodes or line segments) to different locations using predefined criteria.

We discuss these two typical functions in the sections below.

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*, respectively. 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 on 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 5.24.



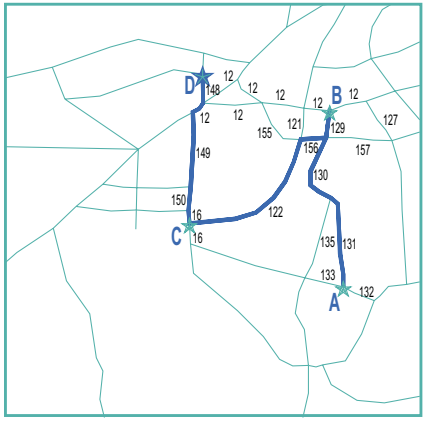
from	onto	cost
a	a	∞
a	b	5
b	a	8
b	b	15
c	a	3
c	b	6

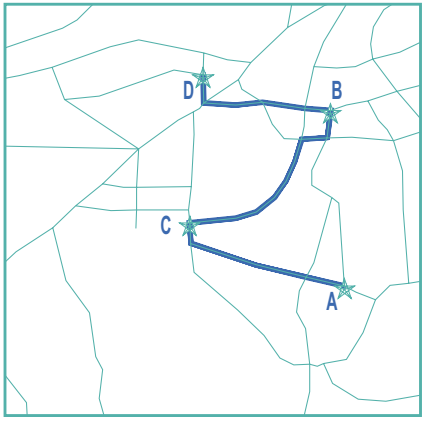
Figure 5.24: Network neighbourhood of node N with associated turning costs at N. Turning at N onto c is prohibited because of direction, so no costs are mentioned for turning onto c. A turning cost of infinity (∞) means that it is also prohibited.

The attentive reader will notice that it is possible to travel on line b in Figure 5.24, then take 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 comes 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 to another node L from M. The turn at M towards node L coming via another line may be prohibitively expensive, whereas turning towards L at M returning to M along b may not be so expensive.

Problems related to optimal path finding are *ordered* optimal path finding and *unordered* optimal path finding. Both have as extra requirement that a number of additional nodes needs 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 5.25. Here, a path is found from node A to node D, visiting 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 GIS provide support for these more complicated path finding problems.

(a) 



(b)

Figure 5.25: 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) both nodes also but in arbitrary order.

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 centre for the network. This may be any type of service: medical treatment, education, water supply. This type of network partitioning is known as a *network allocation problem*.

Another problem is *trace analysis*. Here, one wants to determine that part of the network that is upstream (or downstream) from a given target location. Such problems exist in pollution tracing along river/stream systems, but also in network failure chasing in energy distribution networks.

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. These further 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, school pupil positions, 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 work(wo)men.

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 5.26(a), the green star indicates a primary school and the GIS has been used to assign streets and street segments

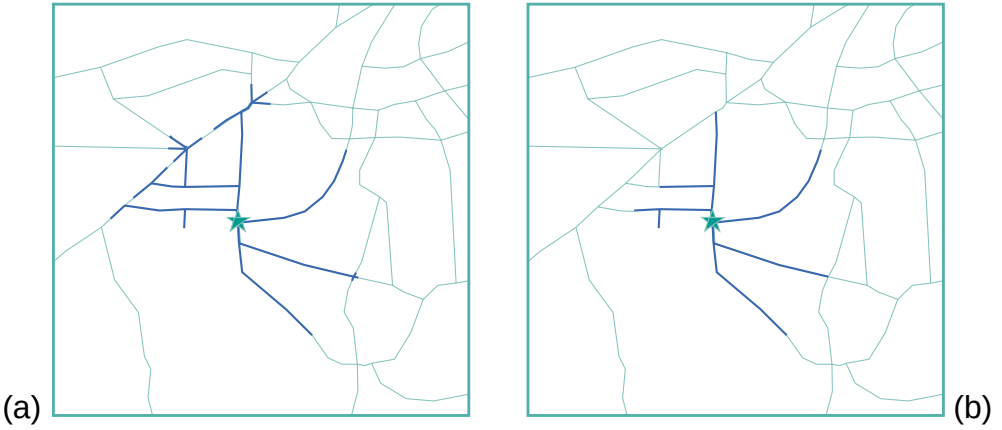


Figure 5.26: Network allocation on a pupil/school assignment problem. In (a), the street segments within 2 km of the school are identified; in (b), the selection of (a) is further restricted to accommodate the school's capacity for the new year.

to the closest school within 2 km distance, along the network. Then, using demographic figures of 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 already selected part of the network was reduced to accommodate precisely the school's pupil capacity for the new year.

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, known as the *trace origin*. For a node or line to be conditionally connected, it 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, and they may involve direction of the path, capacity, length, resource consumption along it, *et cetera*. The condition typically is a logical expression, as we have seen before, for instance:

- 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.

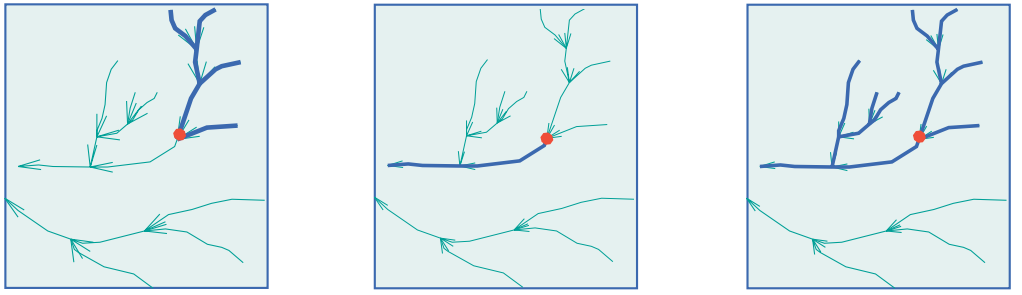


Figure 5.27: Tracing functions on a network: (a) tracing upstream, (b) tracing downstream, (c) tracing without conditions on direction.

In Figure 5.27 our trace origin is indicated in red. In part (a), the tracing conditions were set to trace all the way upstream; in part (b), we traced all the way downstream, and in part (c) we set no conditions on direction of the path, thereby tracing all connected lines from the trace origin. More complicated conditions are certainly possible in tracing.

Summary

In this chapter, we looked at various ways of manipulating spatial data sets, both of the raster and of the vector type. An important distinction is whether our manipulations generate new spatial data sets or not. Throughout the chapter, we have attempted to strike a balance between vector and raster manipulations, trying to give them equal attention, but it is certainly true that some types of manipulations are better accommodated in one, and not so well in the other. But this is not an applicability contest between these two data formats. Usually, one chooses the format to work with on the basis of many more parameters, including source data availability.

A first class of spatial data manipulations does not generate new spatial data, but rather extracts—i.e., 'makes visible'—information from existing data sets. Amongst these are the *measurement* functions. These allow us to determine scalar values such as length, distance, and area size of selected features. Another prominent data extraction type are the *spatial selections*, which allow to selective identify features on the basis of conditions, which may be spatial in character.

The second class of spatial data manipulations does generate new spatial data sets. *Classification* functions come first to mind: they assign a new characteristic value to each feature in a set of (previously selected) features. This then allows to lump features with the same characteristic value together.

Spatial overlay functions go a step further and combine two spatial data sets by location. What is produced as output spatial data set depends on user requirements, and the data format with which one works. Most of the vector spatial overlays are based on polygon/polygon intersection, or polygon/line intersections. In the raster domain, we have seen the powerful tool of raster calculus, which allows all sorts of spatial overlay conditions *and* output expressions, all

based on cell by cell comparisons and computations.

Going beyond spatial overlays are the *neighbourhood* functions. Their principle is not 'equal location comparison' but they instead focus on the definition of the vicinity of one or more features. This is useful for applications that attempt to assess the effect of some phenomenon on its environment. The simplest neighbourhood functions are insensitive to direction, i.e., will deal with all directions equally. Good examples are buffer computations on vector data. More advanced neighbourhood functions take into account local factors of the vicinity, and therefore are sensitive to direction. Since such local factors are more easily represented in raster data, this is then the preferred format. Spread and seek functions are examples.

We finally also looked at special type of spatial data, namely (line) networks, and the functions that are needed on these. *Optimal path finding* is one such functions, useful in routing problems. The use of this function can be constrained or unconstrained. Another function often needed on networks is *network partitioning*: how to assign which parts of the network to which resource locations.

Questions

1. On page 283, we discussed the measurement function of distance between vector features. Draw six diagrams, each of which contains two arbitrary vector features, being either a point, a polyline, or a polygon. Then, indicate the minimal distance, and provide a short description of how this could have been computed.

2. On page 283, we mentioned that two polygons can only intersect when their minimal bounding boxes overlap. Provide a counter-example of the inverted statement, in other words, show that if their minimal bounding boxes overlap, the two polygons may still not intersect (or meet, or have one contained in the other).

3. In Figure 5.11 we provided an example of automatic classification. Rework the example and show what the results would be for three (instead of five) classes, both with equal interval classification and equal frequency classification.

4. A small puzzle on measurements on raster data. Suppose we have a raster with cell area size 15 $\times$ 15 m, and that the raster uses a Cartesian coordinate system. What then is the distance between the cell with 'matrix coordinates' $(12, 113)$ and the cell with coordinates $(84, 556)$? What is the size in reality of the area covered by the minimal bounding box of a line from the first to the second cell?

5. What is the more natural user-controlled classification technique for a raster data set that has not yet been classified?

6. In Figure 5.9, we provided a classification of average household income per ward in the city of Dar es Salaam. Provide a (spatial) interpretation of that figure.

7. Observe that the *equal frequency technique* applied on the raster of Figure 5.11 does not really produce categories with equal frequencies. Explain why this is. Would we expect a better result if our raster had been 5,000 × 5,000 cells?

8. When discussing vector overlay operators, we observed that the one fundamental operator was polygon intersection, and that other operators were expressible in terms of it. The example we gave showed this for polygon overwrite. Draw up a series of sketches that illustrates the procedure. Then, devise a technique of how polygon clipping can be expressed and illustrate this too.

9. Argue why spread computations are much more naturally supported by raster data than by vector data.

10. In Figure 5.22(b), each cell was assigned the minimum total resistance of a path from the source location to that cell. Verify the two values of 14.50 and 14.95 of the top left cells by doing the necessary computations.

11. In Figure 5.23, we illustrated drainage pattern computations on the basis of an elevation raster. Pick two arbitrary cells, and determine how water from those cells will flow through the area described by the raster. Which raster cell can be called the 'water sink' of the area?

Chapter 6

Data visualization

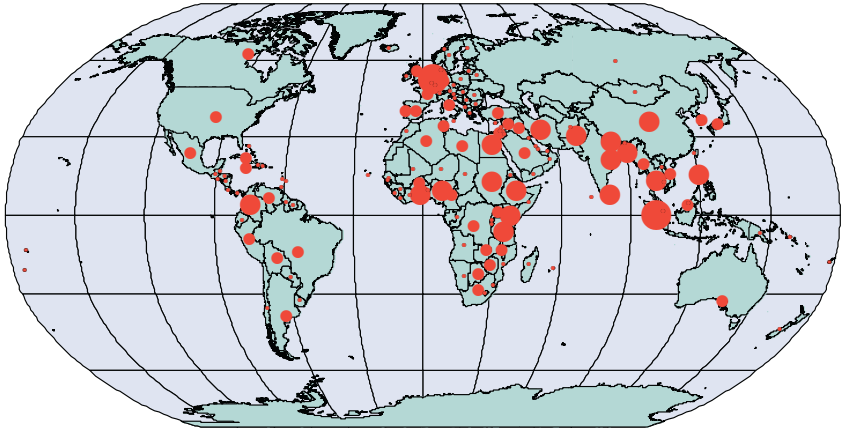


Figure 6.1: Maps and location—"Where did ITC cartography students come from?" Map scale is $1 : 200,000,000$.

6.1 GIS and maps

The relation between maps and GIS is rather intense. Maps can be used as input for a GIS. They can be used to communicate results of GIS operations, and maps are tools while working with GIS to execute and support spatial analysis operations. As soon as a question contains a phrase like "where?" a map can be the most suitable tool to solve the question and provide the answer. "Where do I find Enschede?" and "Where did ITC's students come from?" are both examples. Of course, the answers could be in non-map form like "in the Netherlands" or "from all over the world." These answers could be satisfying. However, it will be clear these answers do not give the full picture. A map would put the answers in a spatial perspective. It could show where in the Netherlands Enschede is to be found and how it is located with respect to Schiphol–Amsterdam airport, where most students arrive. A world map would refine the answer "from all over the world," since it reveals that most students arrive from Africa and Asia,

and only a few come from the Americas, Australia and Europe as can be seen in Figure 6.1.

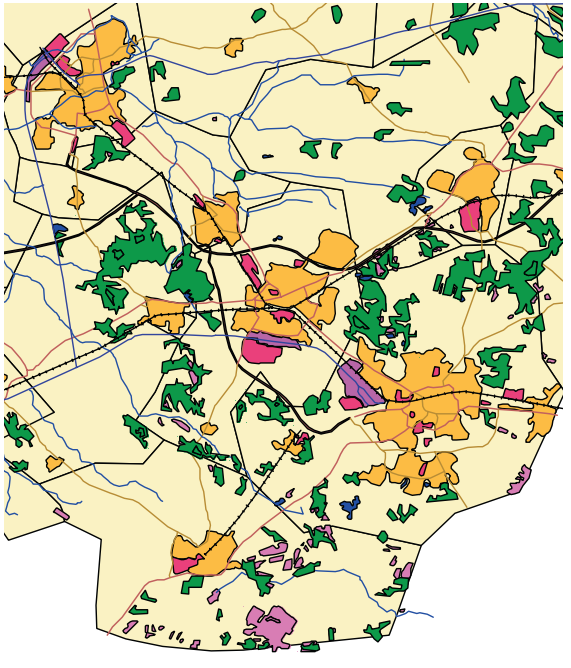


Figure 6.2: Maps and characteristics—"What is the predominant land use in southeast Twente?"

As soon as the location of geographic objects ("where?") is involved a map is useful. However, maps can do more then just providing information on location. They can also inform about the thematic attributes of the geographic objects located in the map. An example would be "What is the predominant land use in southeast Twente?" The answer could, again, just be verbal and state "Urban." However, such an answer does not reveal patterns. In Figure 6.2, a dominant northwest-southeast urban buffer can be clearly distinguished. Maps can an-

swer 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 will probably be satisfactory to most people. However, it might be interesting to see how this changed over the years. A set of maps could provide the answer as demonstrated in Figure 6.3. Summarizing, maps can deal with questions/answers related to the basic components of spatial or geographic data: location (geometry), characteristics (thematic attributes) and time, and their combination.

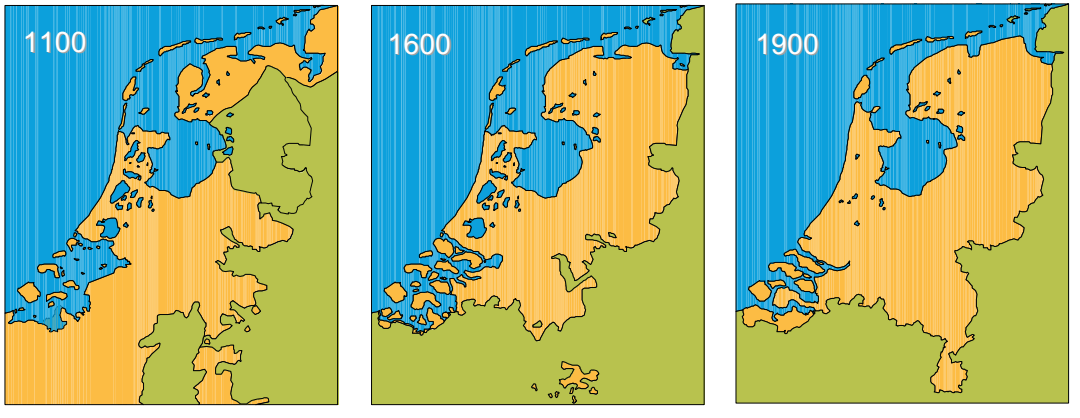
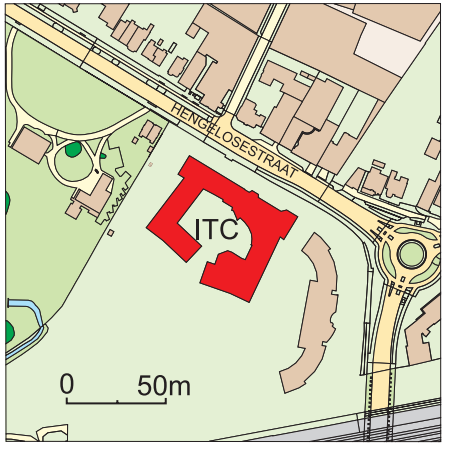


Figure 6.3: Maps and time—"When did the Netherlands have its longest coastline?"

As such, maps are the most efficient and effective means to transfer 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 in 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, and as such allow for more complex queries.

Looking at the maps in this paragraph's illustrations 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 it puts, when well-designed, the remaining information in a clear perspective. The map in Figure 6.1 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.

(a)  

(b)

Figure 6.4: Comparing an aerial photograph (a) and a map (b). Source: Figure 5–1 in [36].

This characteristic is well illustrated when one puts the map next to an aerial photograph or satellite image of the same area. Products like these give all information observed by the capture devices used. Figure 6.4 shows an aerial photograph of the ITC building and a map of the same area. The photographs show all objects visible, including parked cars, small temporary buildings *et cetera*. From the photograph, it becomes clear that the weather as well as the time of the day

influenced its contents: the shadow to the north of the buildings obscures other information. The map only gives the outlines of buildings and the streets in the surroundings. It is easier to interpret because of selection/omission and classification. The symbolization chosen highlights our building. Additional information, not available in the photograph, has been added, such as the name of the major street: Hengelosestraat. Other non-visible data, like cadastral boundaries or even the sewerage system, could have been added in the same way. However, it 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.

Apart from contents, there is a relationship between the effectiveness of a map for a given purpose and the 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 6.1 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 6.4 displaying the surroundings of the ITC-building is an example. The world map in Figure 6.1 is a *small-scale map*. Scale indications on maps can be given verbally like '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 $2,000$ km) in reality), or by a graphic representation like a scale bar as given in the map in Figure 6.4(b). The advantage of using scale bars in digital environments is that its length changes also when the map zoomed in, or enlarged before printing.[1] Sometimes it is necessary to convert maps from one scale to another, but this may lead to problems of (cartographic) *generalization*.

Having discussed several characteristics of maps it is now necessary to pro-

[1] And this explains why many of the maps in this book do not show a map scale.

vide a definition. Board [8] 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 holds three key words. The geographic reality represents the object of study, 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 (parts of) the Earth's surface on a plane.

Traditionally, maps are divided in *topographic* 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., railroads and roads), land use (e.g., vegetation and built-up area), relief, hydrology, geographic names and a reference grid. Figure 6.5 shows a small scale topographic map 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 6.6(a), showing population density in Overijssel, is an example of the first and the map in Figure 6.6(b), displaying the province's drainage areas, is an example of the second. As can be noted, both thematic maps also contain information found in a topographic map, 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 with drainage areas should have added rivers and canals, while adding relief would make sense as

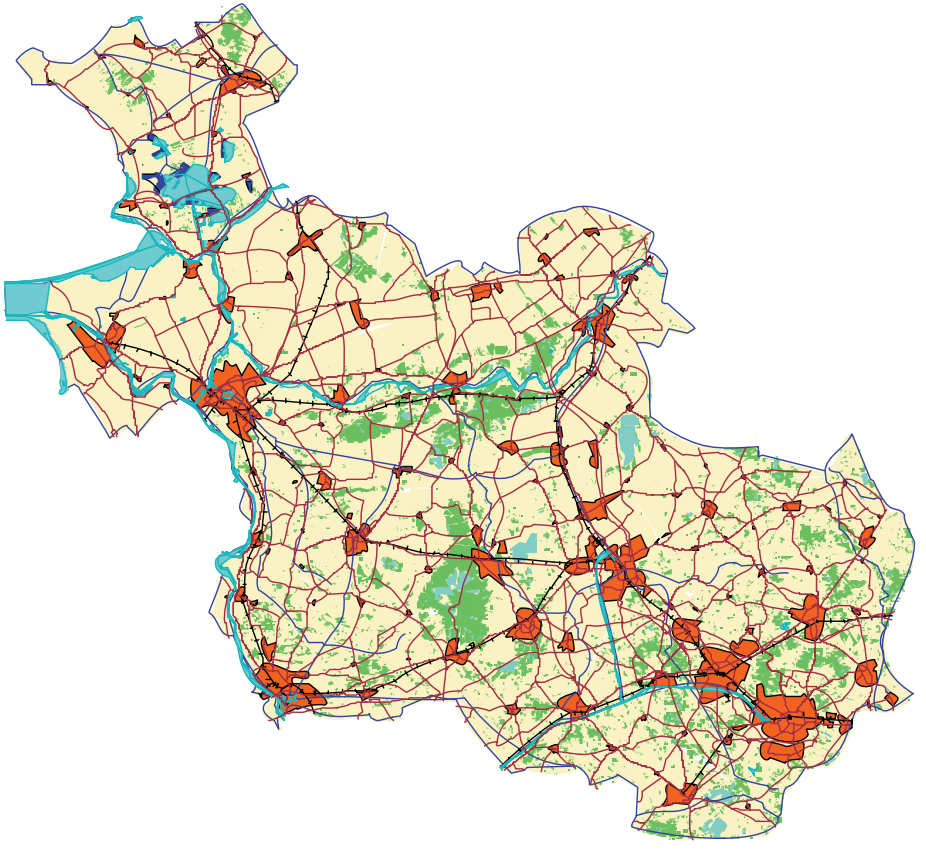


Figure 6.5: A topographic map of the province of Overijssel. Geographic names and a reference grid have been omitted for reasons of clarity.

well. Today's digital environment has diminished the distinction between topographic and thematic maps. Often, both topographic and thematic maps are stored in the 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.

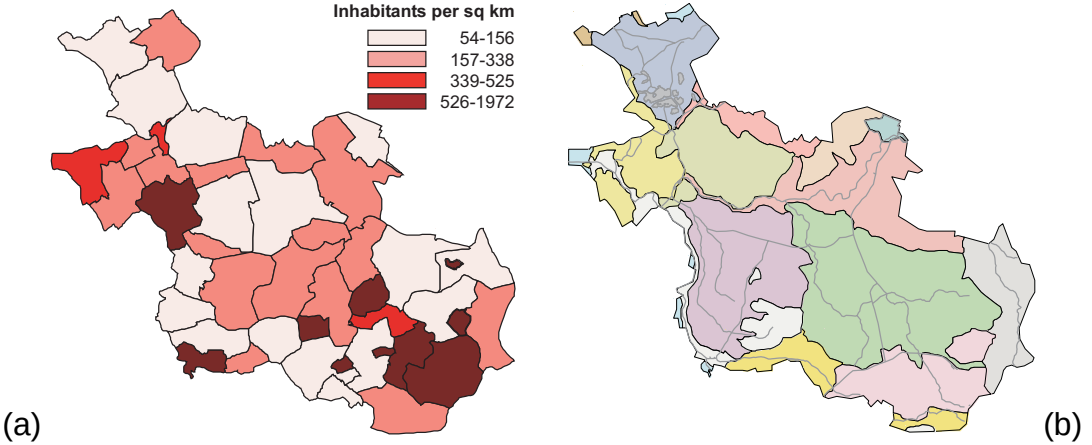


Figure 6.6: Thematic maps: (a) socio-economic thematic map, showing population density of province of Overijssel (higher densities in darker tints); (b) physical thematic map, showing watershed areas of Overijssel.

The design of topographic maps is mostly based on conventions, of which some date back to centuries ago. Examples are water in blue, forests in green, major roads in red, urban areas in black, *et cetera*. The design of thematic maps, however, should be based on a set of cartographic rules, also called *cartographic grammar*, which will be explained in Section 6.4 and 6.5 (but see also [37]).

Nowadays, maps are often produced through a GIS. If one wants to use a GIS to tackle a particular geo-problem, this often involves the combination and integration of many different data sets. For instance, if one wants to quantify land use changes, two data sets from different periods can be combined with an overlay operation. The result of such a spatial analysis can be a spatial data layer

from which a map can be produced to show the differences. The parameters used during the operation are based on computation 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. From this perspective, maps are no longer only the final product they used to be. They can be created just to see which data are available in the spatial database, or to show intermediate results during spatial analysis, and of course to present the final outcome.

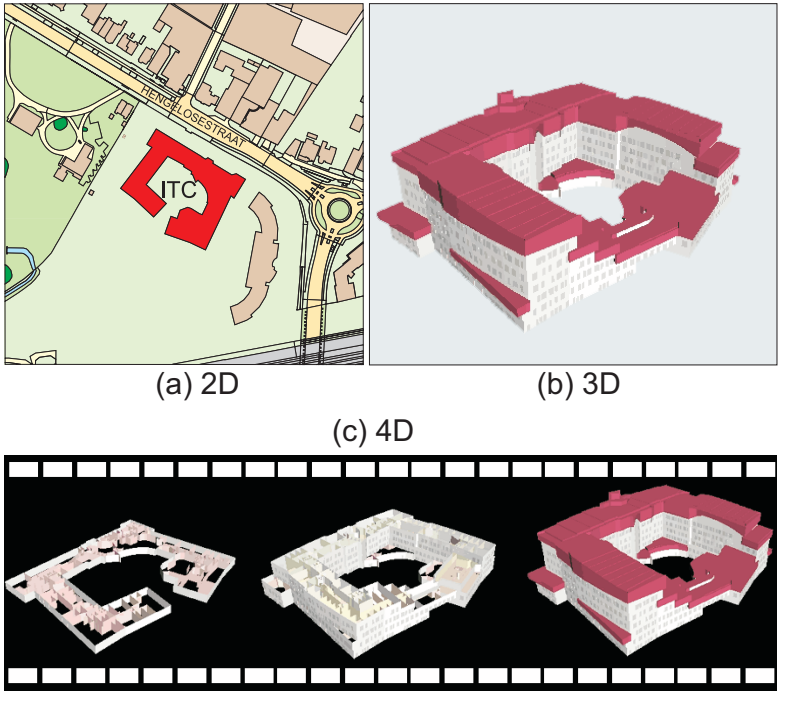


Figure 6.7: The dimensions of spatial data: (a) 2D, (b) 3D, (c) 3D with time.

The users of GIS 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 6.7 shows the dimensionality of geographic objects and their graphic representation. Part (a) provides a map of the ITC building and its surroundings, while part (b) shows a three-dimensional view of the building. Figure 6.7(c) shows the effect of change, as two moments in time during the construction of the building.

6.2 The visualization process

The characteristic of maps and their function in relation to the spatial data handling process was 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. These 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, the production and use of maps, depending on the application (see Figure 6.8).

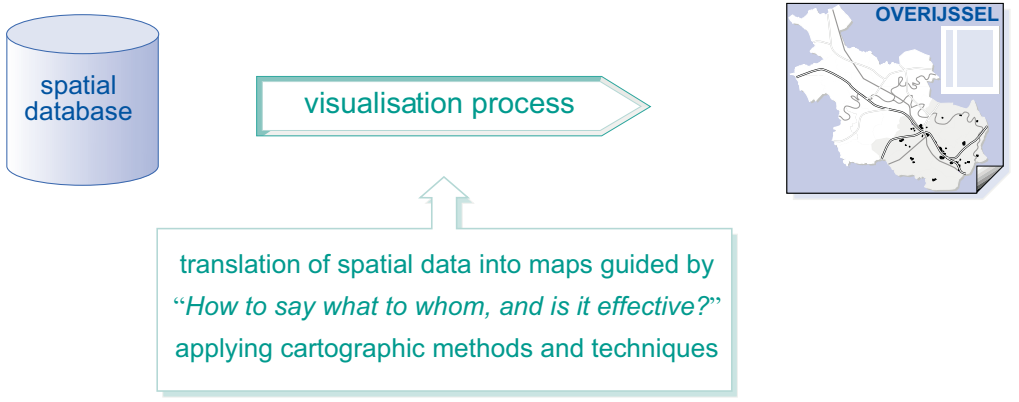


Figure 6.8: The cartographic visualization process. Source: Figure 2–1 in [36].

The producer of these visual products may be a professional cartographer, but may also be a discipline expert mapping, for instance, vegetation stands using remote sensing images, or 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 in 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. This will be elaborated upon in the following sections.

In the past, the cartographer was 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 by 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 on its efficiency. To the proposition "How do I say what to whom" we have to add "and is it effective?" Based on feedback from map users, we can decide whether the map needs improvement. In particular, with all the modern visualization options available, such as animated maps, multimedia and virtual reality, it remains necessary to test cartographic products on their effectiveness.

The visualization process is always influenced by several factors, as can be illustrated by just looking at the content of a spatial database:

- Are we dealing with large- or small-scale data? 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 traditionally resulted in different design approaches as was explained in the previous section.
- More important for the design is the question of whether the data to be represented are of a quantitative or qualitative nature.

We should understand that the impact of these factors may become even bigger 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 to classify the data or to smoothen 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 in blue, lowlands in green and mountains in brown. The efficiency of these tools will partly depend on the above-mentioned factors, and partly on what we are used to.

6.3 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, meaning, to inform the map user about location and 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 [55, 37].

During the last decades, many others have become involved in making maps. The widespread use of GIS has increased the number of maps tremendously [42]. Even the spreadsheet software used commonly in office 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 dealing with spatial data. 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 process.

This thinking process is accelerated by the continued developments in hard- and software. These went along with changing scientific and societal needs for georeferenced data and, as such, for maps. New media like CD-ROMs and the WWW allow *dynamic presentation* and also *user interaction*. 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 other sectors. We lack the tools for user-friendly queries and retrieval when studying the massive amount of data produced by sensors, which is now available via the WWW. A new branch of science is currently evolving to solve this problem of abundance. In the geo-

disciplines, it is called *visual spatial data mining*.

The developments have given the word *visualization* an enhanced meaning. According to the dictionary, it means 'to make visible' and 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 it [44], and this has an important impact on cartography as well. It 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 function to the map. As well as communication, it will prompt thinking and decision-making.

Developments in scientific visualization stimulated DiBiase [18] to define a model for map-based scientific visualization, also known as geovisualization. It covers both the presentation and exploration functions of the map (see Figure 6.9). Presentation is described as 'public visual communication' since it concerns maps aimed at a wide audience. Exploration is defined as 'private visual thinking' because it is often 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 use. Exploration, however, often involves a discipline expert 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, also here the problem of translation of

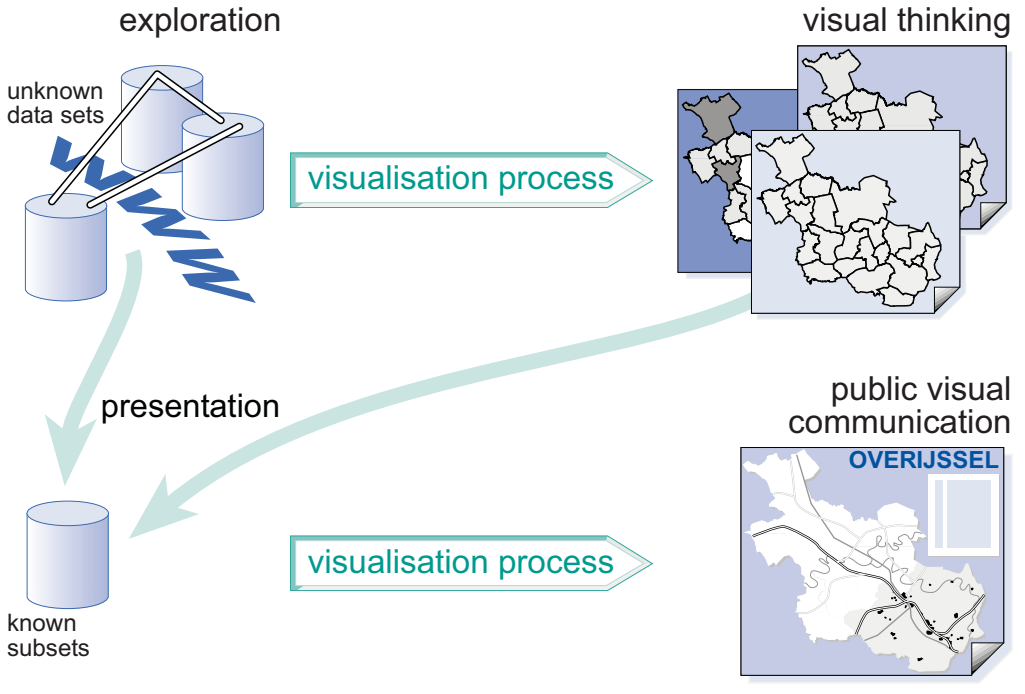


Figure 6.9: Visual thinking and visual communication. Source: Figure 2–2 in [36].

spatial data into cartographic symbols needs to be solved.

The above trends have all to do with what has been called the 'democratization of cartography' by Morrison [47]. He explains it as

> "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 working with unknown patterns in data. However, what is unknown for one is not necessarily unknown to others. For instance, browsing in Microsoft's *Encarta World Atlas* CD-ROM is an exploration for most of us because of its wealth of information. With products like these, such exploration takes place within boundaries set by the producers. Cartographic knowledge is incorporated in the program, resulting in pre-designed maps. Some users may feel this to be a constraint, but those same users will no longer feel constrained as soon as they follow the web links attached to this electronic atlas. It shows that the environment, the data and the users influence one's view of what exploration entails.

To create a map about a topic means that one selects the relevant geographic phenomena according to some model, and converts these into meaningful symbols for the map. Paper maps (in the past) had a dual function. They acted as a database of the objects selected from reality, and communicated information about these 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 it. The communicative 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 that are created in the realm of presentation cartography (Figure 6.9), it is important to adhere to the cartographic design rules. This is to guarantee that they are easily understood by the map users.

How does this communication process work? Figure 6.10 forms an illustration. It starts with information to be mapped (the 'What' from the sentence).

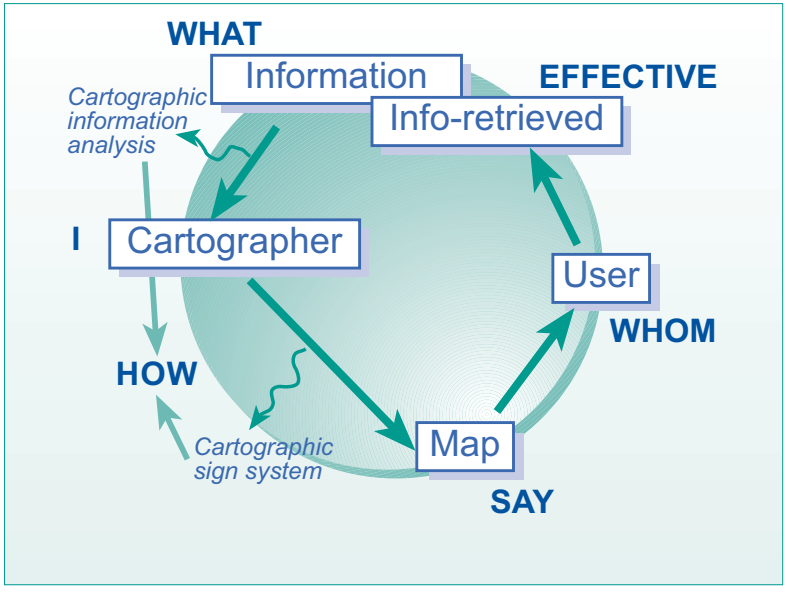


Figure 6.10: The cartographic communication process, based on "How do I say what to whom, and is it effective?" Source: Figure 5–5 in [36].

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 this. Based on this knowledge, the cartographer can choose the correct symbols to represent the information in the map. S/he has a whole toolbox of visual variables available to match symbols with the nature of the data. For the rules, we refer to Section 6.4.

In 1967, the French cartographer Bertin developed the basic concepts of the theory of map design, with his publication *Sémiologie Graphique* [6]. He provided guidelines for making good maps. If ten professional cartographers were given the same mapping task, and each would apply Bertin's rules, this would still

result in ten different maps. For instance, if the guidelines dictate the use of colour, it is not stated which colour should be used. Still, all ten maps could be of good quality.

Returning to the scheme, the map (the 'say' in the sentence) is read by the map users (the 'whom' from the sentence). They extract some information from the map, represented by the box entitled 'retrieved information'. From the figure it becomes clear that the boxes with 'information' and 'retrieved Information' 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. There may be several causes. Possibly, the original information was partly lost or additional information has been added during the process. Loss of information could be deliberately caused by the cartographer, with the aim to emphasize remaining information. Another possibility is that the map user did not understand the map fully. Information gained during the communication process could be due to the cartographer, who added extra information to strengthen the already available information. It is also possible that the map user has some prior knowledge on the topic or area, which allows the user to combine this prior knowledge with the knowledge retrieved from the map.

6.4 The cartographic toolbox

6.4.1 What kind of data do I have?

To find the proper symbology for a map one has to execute a cartographic data analysis. The core of this analysis process is to access the characteristics of the data to find out how they can be visualized, so that the map user properly interprets them. 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 geomorphology the title will be *Geomorphology of ...* Secondly, the individual component(s), such as those that relate to the origin of the land forms, should be accessed and their nature described. Later, these components should be visible in the map legend. Analysis of the components is done by determining their nature.

Data will be of a *qualitative* or *quantitative* nature. The first type of data is also called *nominal* data. Nominal data exist of discrete, named values without a natural order amongst the values. Examples are the different languages (e.g., English, Swahili, Dutch), the different soil types (e.g., sand, clay, peat) or the 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. Basic geographic units are homogeneous areas associated with a single soil type, recognized 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 exists no absolute zero on the scale. Temperature is an example: 40 °C is not twice as warm as 20 °C, and 0 °C is not an absolute zero. Quantitative data with a ratio scale have a known absolute zero. An example is income: someone earning $100 earns twice as much as someone with an income of $50. In the 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 an ordinal scale, based on hierarchies. For instance, one knows that one 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 6.1.

Measurement scale	*Nature of data*
Nominal	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

Table 6.1: Differences in the nature of data and their measurement scales

6.4.2 How can I map my data?

The contents of a map, irrespective of the medium on which it is displayed, can be classified in different basic categories. A map image consists of point symbols, line symbols, area symbols, and text. The symbols' appearance can vary depending on their nature. Most maps in this book show symbols in different size, shape and colour. Points can represent individual objects such as the location of shops or can refer to values that are representative for an administrative area. Lines can vary in colour to show the difference between administrative boundaries and rivers, or vary in shape to show the difference between railroads and roads. Areas follow the same principles: difference in colour distinguishes between different vegetation stands.

Although the variations are only limited by fantasy they can be grouped together in a few categories.

Bertin [6] distinguished six categories, which he called the *visual variables* and which may be applied to point, line and area symbols. They are

- *size*,
- *(lightness) value*,
- *texture*,
- *colour*,
- *orientation* and
- *shape*.

These visual variables can be used to make one symbol different from another. In doing this, map makers in principle have free choice, provided they do not violate the rules of cartographic grammar. They do not have that choice when deciding where to locate the symbol in the map. The symbol should be located where features belong. Visual variables stimulate the map user's perception in different ways. What is perceived depends on the human capacity to see what belongs together (e.g., all red symbols represent danger), to see order (e.g., the population density varies from low to high—represented by light and dark colour tints, respectively), to perceive quantities (e.g., symbols changing in size with small symbols for small amounts), or to get an instant overview of the mapped theme. The next section will discuss some typical mapping problems and demonstrate the above.

6.5 How to map . . . ?

The subsections in this *How to map . . .* section deal with characteristic mapping problems. We first describe a problem and 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 wrongly designed but are commonly found.

6.5.1 How to map qualitative data

If one, after a long fieldwork period, has finally delineated the boundaries of a province's watersheds, one likely is interested in 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, and none should stand out above the others.

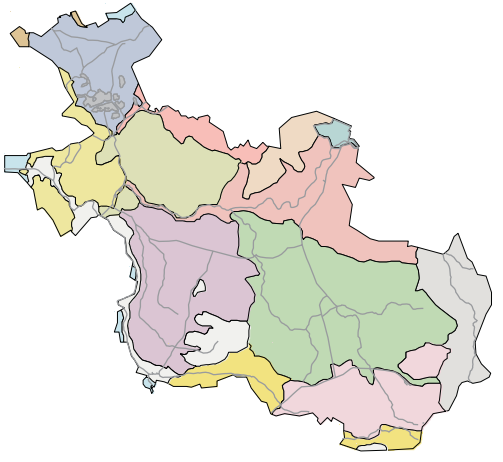
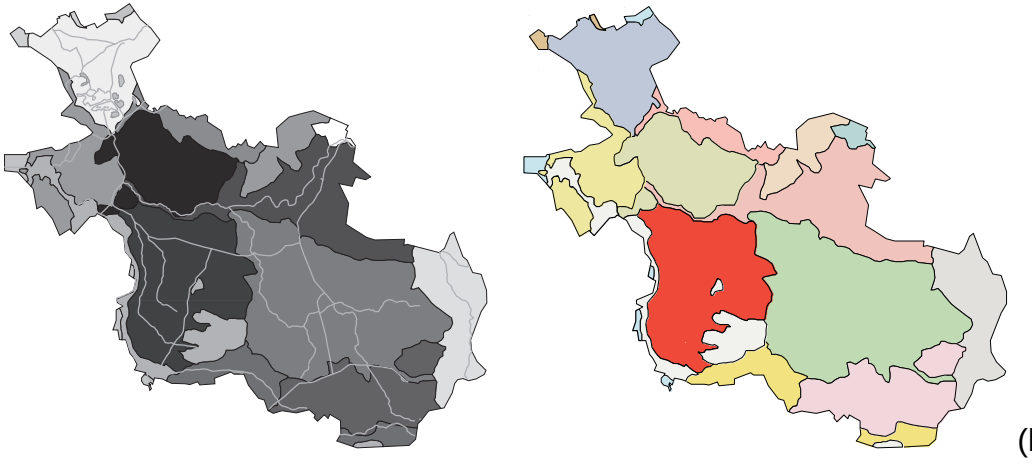


Figure 6.11: A good example of mapping qualitative data

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 6.11 gives an example of a correct map. The readability is influenced by the number of displayed geographic units. In this example, there are about 15. When this number is over one hundred, the map, at the scale displayed here, will become too clut-

tered. The map can also be made with different black and white patterns—as an application of the visual variable shape—to distinguish between the watersheds. The amount of geographic units that can be displayed is even more critical then.

Figure 6.12 shows two examples of how *not* to create such a map. In (a), several tints of black are used—as application of the visual variable lightness value. Looking at the map may cause perceptual confusion since the map image suggests differences in importance that are not there. In Figure 6.12(b), colours are used instead. However, where most watersheds are represented in pastel tints, one of them stands out by its bright colour. This gives the map an unbalanced look. The viewer's eye will be distracted by the bright colours, resulting in an unjustified weaker attention for other areas.



(a) (b)

Figure 6.12: Two examples of wrongly designed qualitative maps: (a) misuse of tints of black; (b) misuse of bright colours

6.5.2 How to map quantitative data

When, after executing a census, one would for instance like to create a map with the number of people living in each municipality, one deals 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 geographic distribution of the phenomenon. To reach this objective, the symbols used should have quantitative perception properties. Symbols varying in size fulfil this demand. Figure 6.13 shows the final map for the province of Overijssel.

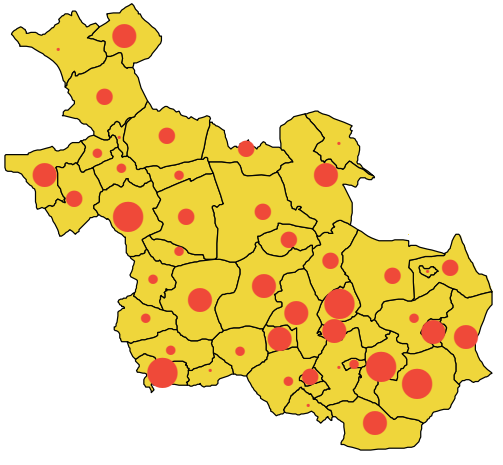


Figure 6.13: Mapping absolute quantitative data

That it is easy to make errors can be seen in Figure 6.14. In 6.14(a), different tints of green 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 ge-

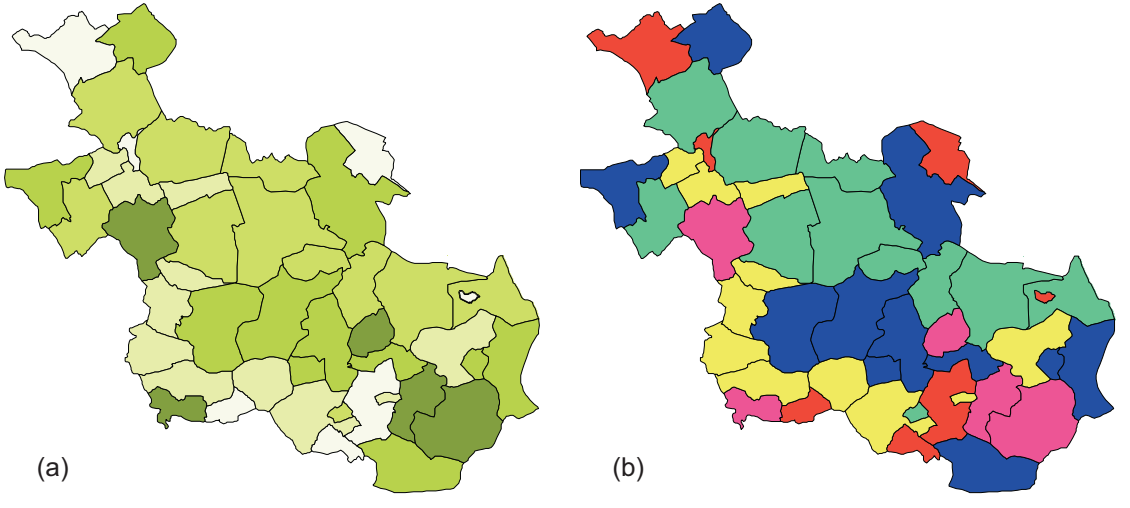


Figure 6.14: Poorly designed maps displaying absolute quantitative data: (a) wrong use of green tints for absolute population figures; (b) incorrect use of colour

ographic unit 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 argument is that the population is not necessarily homogeneously distributed within the geographic units. Colour has also been misused in Figure 6.14(b). The applied four-colour scheme makes it is impossible to say whether red represents more populated areas than blue. It is impossible to instantaneously answer a question like "Where do most people in Overijssel live?"

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 then deal with *relative quantitative data*. The numbers now have a clear relation with the area they represent. The geographic unit will again be

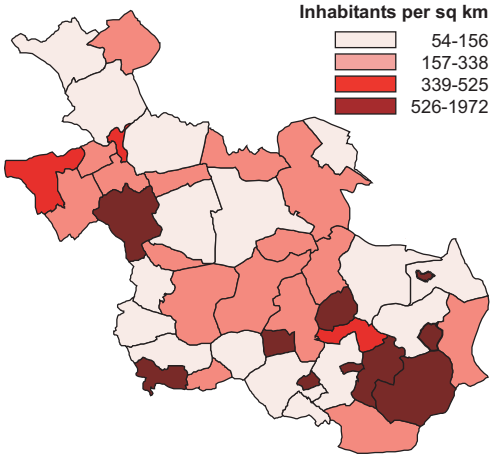


Figure 6.15: Mapping relative quantitative data

municipality. Aim of the map is to give an overview of the distribution of the population density. In the map of Figure 6.15, value has been used to display the density from low (light tints) to high (dark tints). The map reader will automatically and in a glance associate the dark colours with high density and the light values with low density. Figure 6.16(a) 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 through quite some trouble to find out where in the province the high-density areas can be found. Why should mid-red represent areas with a higher population density then dark-red? In Figure 6.16(b) colour has been used in combination with lightness value. The first impression of the map reader would be to think the brown areas represent the areas with the highest density. A closer look at a legend would tell that this is not the case, and that those areas are represented by another colour that did not speak for itself.

If one really studies the badly designed maps carefully, the information can

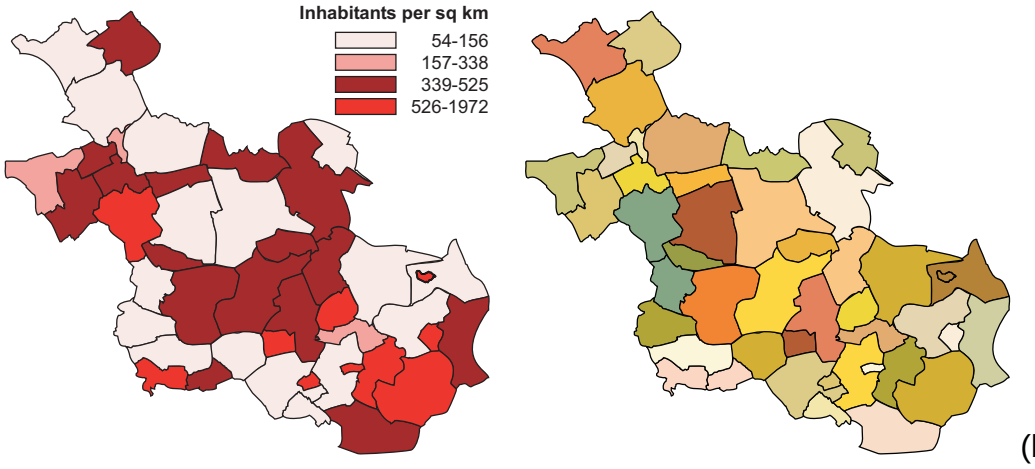


Figure 6.16: Badly designed maps representing relative quantitative data: (a) lightness values used out of sequence; (b) colour should not be used

be derived, in one way or another, but it would take quite some effort. Proper application of cartographic guidelines will guarantee that this will go much more smoothly (e.g., faster and with less chance of misunderstanding).

6.5.3 How to map the terrain elevation

Terrain elevation can be mapped using different methods. Often, one will have collected an elevation data set for individual points like 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 lines connect points of equal elevation, is generally used. To visually improve the information content of such a map, usually only at small scales, the space between the contour lines can be filled with colours following a convention: green for low and brown for high elevation areas. Even more advanced is the addition of *shaded relief*. This will improve the impression of the three-dimensional relief (see Figure 6.17).

The shaded relief map also uses the full three-dimensional information to create shading effects. This map, represented on a two-dimensional surface, can be floated in three-dimensional space to give it a real three-dimensional appearance, as shown in Figure 6.17(d). Looking at such a representation one can immediately imagine that it will not always be effective. Certain objects in the map will easily disappear behind other objects. Interactive functions to manipulate the map in three-dimensional space so as to look behind some objects are required. 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 others require large-scale resolution. 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. 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 sparsely available.

Thematic data can also be viewed in three dimensions. This may result in

dramatic images, which will be long remembered by the map user. Figure 6.18 shows the absolute population figures of Overijssel in three dimensions. Instead of a proportional symbol that depicts the number of people living in a municipality (as we did in Figure 6.13) the height at a municipality now indicates total population. Since data in two-dimensional maps are often classified in a few categories only, the relations among the geographic objects are easier to understand. The image clearly shows that Enschede (the large column in the lower right) is by far the biggest town.

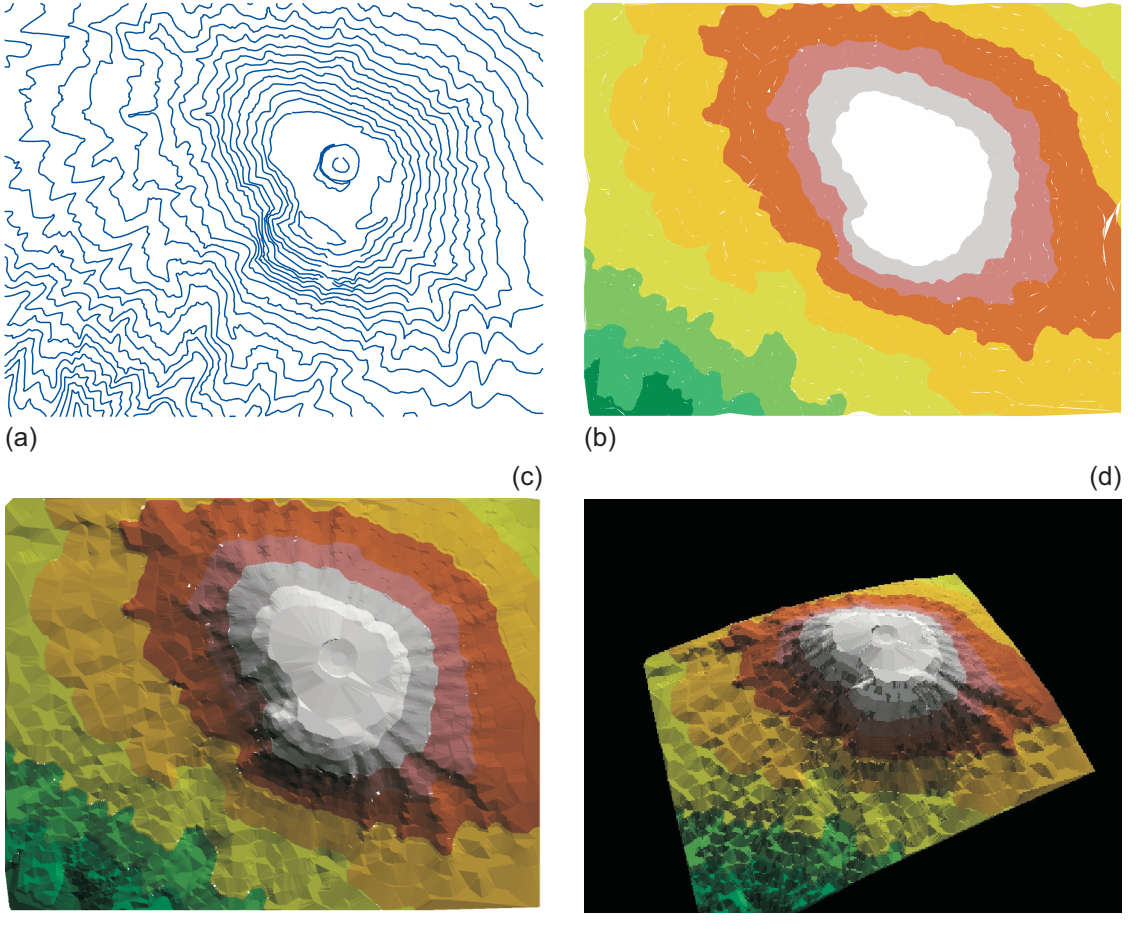


Figure 6.17: visualization of terrain elevation: (a) contour map; (b) map with layer tints; (c) shaded relief map; (d) 3D view of the terrain

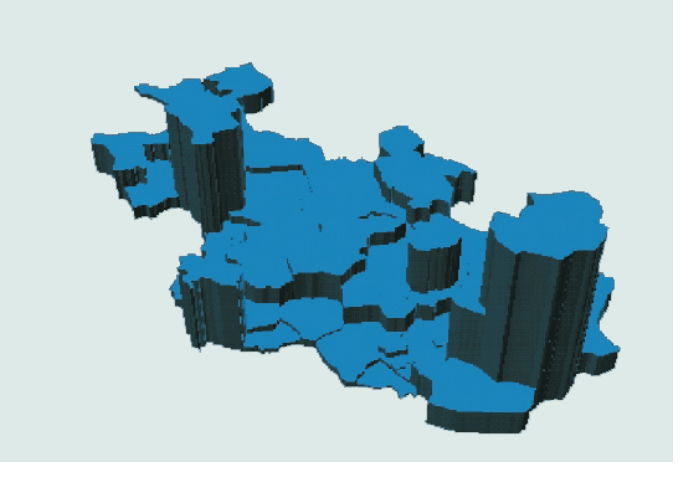


Figure 6.18: Quantitative data visualized in three dimensions

6.5.4 How to map time series

Advances in spatial data handling have not only made the third dimension part of daily GIS routines. Nowadays, the manipulation of time-dependent data is also part of these routines. This has been caused by the increasing availability of data captured at different periods in time. Next to this data abundance, the GIS community wants to analyse changes caused by real world processes. To that end, single time slice data are no longer sufficient, and the visualization of these processes cannot be supported with static paper maps only.

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 6.3, the location of Europe's national boundaries, or the position of weather fronts. The changes of a parcel's owner or changes in road traffic intensity are examples of changing attributes. Urban growth is a combination of both. The urban boundaries expand and simultaneously the land use shifts from rural to urban. If maps have to represent events like these they should be suggestive of such change.

This implies the use of symbols that are perceived as representing change. Examples of such symbols are arrows that have an origin and a destination. They are used to show movement and their size can be an indication of the magnitude of change. Also, specific point symbols such as 'crossed swords' (battle) or 'lightning' (riots) can be used to represent dynamics. Another alternative is the use of 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 6.19(a)).

It is possible to distinguish between three temporal cartographic techniques (see Figure 6.19):

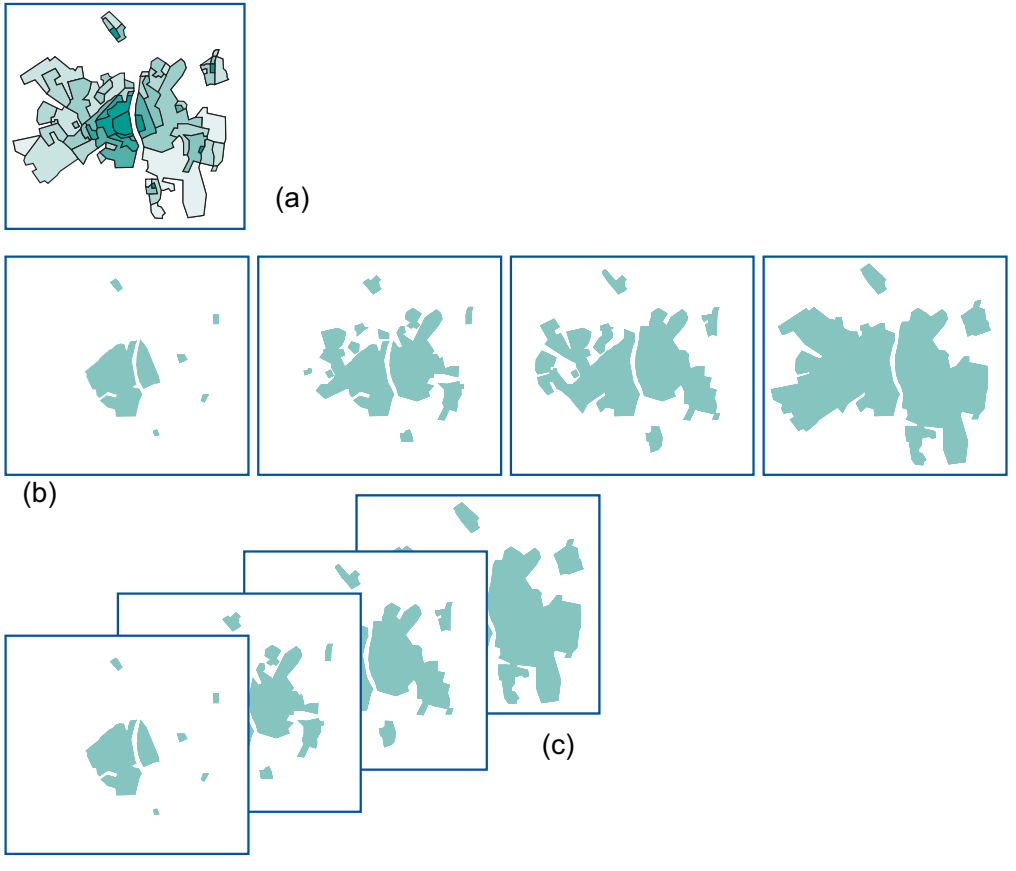


Figure 6.19: Mapping change; example of the urban growth of the city of Maastricht, The Netherlands: (a) single map, in which tints represent age of the built-up area; (b) series of maps; (c) (simulation of an) animation.

Single static map Specific graphic variables and symbols are used to indicate change or to represent an event. Figure 6.19(a) applies colour tints to represent the age of the 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, which the user has to follow, to perceive the temporal variation. The number of images is, however, limited since it is difficult for the human eye to follow long series of maps (Figure 6.19(b));

Animated map Change is perceived to happen in a single image by displaying several snapshots after each other just like a video cut with successive frames. The difference with the series of maps is that the variation is deduced not from a spatial sequence but from real 'change' in the image itself (Figure 6.19(c)).

For the user of a cartographic animation, it is important to have tools available that allow for interaction while viewing the animation. Seeing the animation play will often leave users with many questions about what they have seen. Just replaying the animation is not sufficient to answer questions like "What was the position of the coastline in the north 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 'slow motion'. More options have to be added, such as a possibility to directly go to a certain frame based on a task like: 'Go to 1850'.

6.6 Map cosmetics

Most maps in this chapter are correct from a cartographic grammar perspective. However, many of them lack the information needed to be fully understood. Each map should have, next to the map image, a *title*, informing the user about 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 projection used, and some bibliographic data. The bibliographic data should give the user an idea when the map was created, how old the data used are, who has created the map and even what tools were used. All this information allows the user to obtain an impression of the quality of the map. This information is comparable with metadata describing the contents of a database. Figure 6.20 illustrates these map elements. On paper maps, these elements have to appear next to the map face itself. Maps presented on screen often go 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 title are often available on demand as well.

The map in Figure 6.20 is one of the first in this chapter that has text included. Figure 6.21 is another example. Text is used to transfer information in addition to the symbols used. This can be done by the application of the visual variables to the text as well. In Figure 6.21 more variation can be found. Italics—*cf.* the visual variable of orientation—have been used for building names to distinguish them from road names. The text should also be placed in a proper position with respect to the object it refers to.

Maps constructed via the basic cartographic guidelines are not necessarily appealing maps. Although well-constructed, they might still look sterile. The

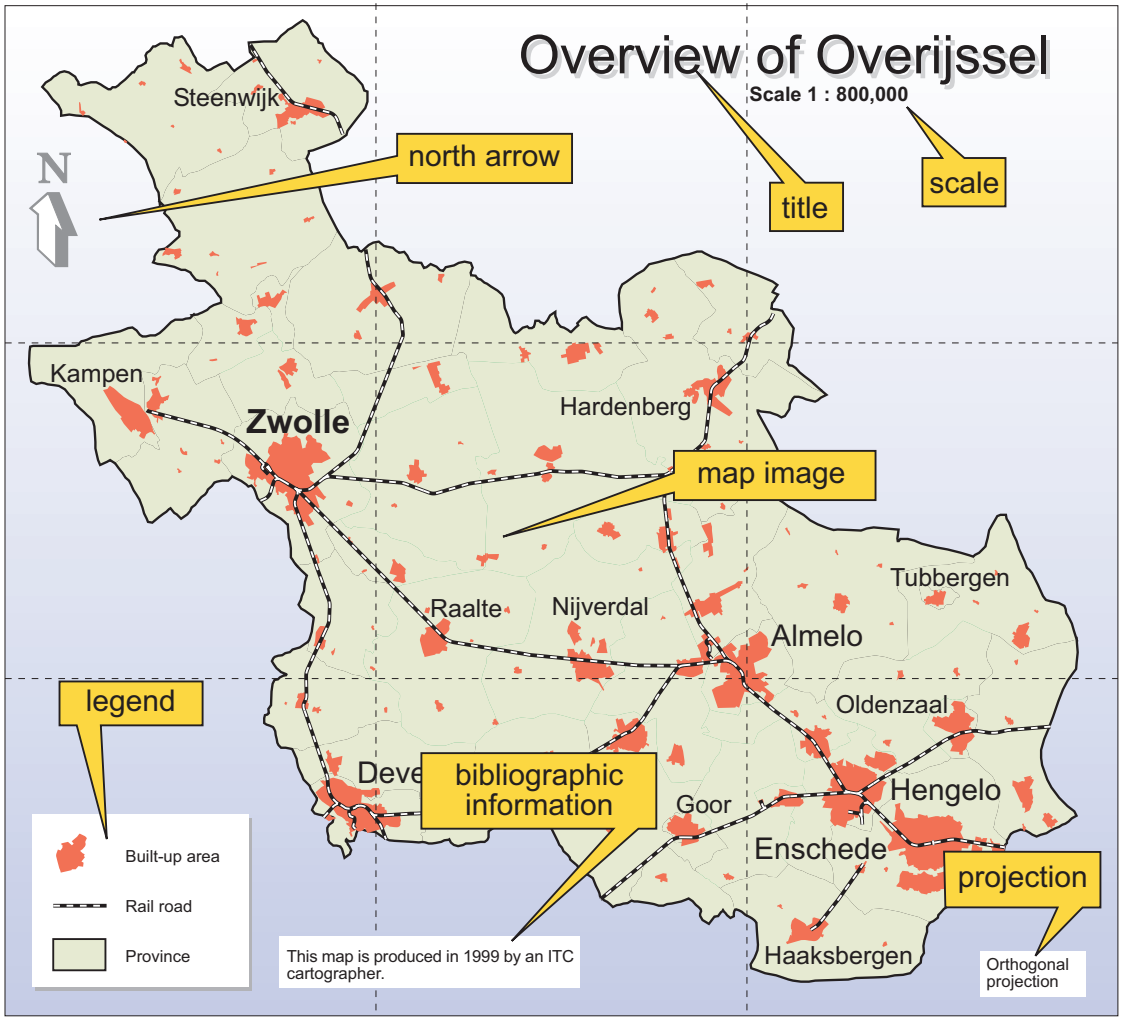


Figure 6.20: The paper map and its (marginal) information. Source: Figure 5–10 in [36].

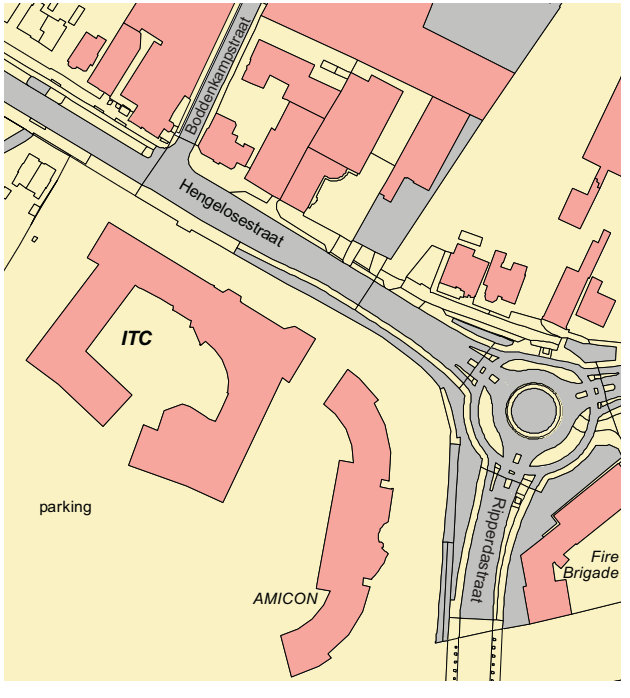


Figure 6.21: Text in the map

design aspect of creating appealing maps has to be included in the visualization process as well. 'Appealing' does not only mean having nice colours. One of the keywords here is *contrast*. Contrast will increase the communicative role of the map since it creates a hierarchy in the map contents, assuming that not all information has equal importance. This design trick is known as *visual hierarchy* or the figure-ground relation. The need for visual hierarchy in a map is best understood when looking at the map in Figure 6.22(a), 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 patches in the map) followed by other buildings, with the road on a lower level and the parcels at the lowest level.

Figure 6.22: Visual hierarchy and the location of the ITC building: (a) hierarchy not applied; (b) hierarchy applied

6.7 Map output

The map design will not only depend on 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") but also on the output medium. Traditionally, maps were produced on paper, and many still are.

Currently, most maps are presented on screen, for a quick view, for an internal presentation or for presentation on the WWW. Compared to maps on paper, on-screen maps have to be smaller, and therefore its contents should be carefully selected. This might seem a disadvantage, but presenting maps on-screen offers very interesting alternatives. In one of the previous paragraphs, we discussed that the legend only needs to be a mouse click away. A mouse click could also open the link to a database, and reveal much more information than a paper map could ever offer. Links to other than tabular or map data could be made available. Maps and multimedia (sound, video, animation) become one, especially in an environment such as the WWW.

On-screen maps should use the opportunities for interaction and dynamics. Blinking and moving map symbols can now be applied to enhance the message of the map. Multimedia allows for interactive integration of sound, animation, text and (video) images. Some of today's electronic atlases, such as the *Encarta Worldatlas* are good examples of how multimedia elements can be integrated with the map. Pointing to a country on a world map starts 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 one of the latest media to present and disseminate spatial data, especially in combination with multimedia elements. In this process, the map plays a key role, and has multiple functions. Maps can play their

traditional role, for instance to provide insight in spatial patterns and relations. But because of the nature of the WWW, the 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, somewhere in cyberspace. Maps can also be used as previews of spatial data products to be acquired.

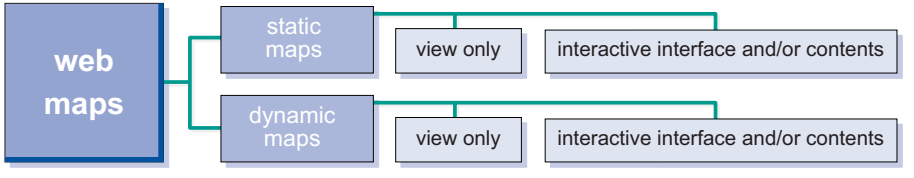


Figure 6.23: Classification of maps on the WWW. Source: Figure 1–2 in [36].

See also Division of Cartography's website on maps on the web

How can maps be used on the WWW? We can distinguish several methods that differ in terms of necessary technical skills from both the user's and provider's perspective. The overview given here (see Figure 6.23) can only be a current state of affairs, since developments on the WWW are tremendously fast. An important distinction is the one between static and dynamic maps.

Most *static maps* on the web are view-only. Many organizations, such as map libraries or tourist information providers, make their maps available in this way. This form of presentation can be very useful, for instance, to make historical maps more widely accessible. Static, view-only maps can also serve to give web surfers a preview of the products that are available from organizations, such as National Mapping Agencies.

When static maps 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 to serve as an interface to spatial data. Clicking on geographic objects may lead the user to quantitative data, photographs, sound or video or other information sources on the Web.

The user may also interactively determine the contents of the map, by choosing data layers, and even the visualization parameters, by choosing symbology and colours. Dynamic maps are about change; change in one or more of the spatial data components. On the WWW, several options to play animations are available. The so-called animated-GIF can be seen as a view-only version of a dynamic map. A sequence of bitmaps, each representing a frame of an animation, are positioned one after another, and the WWW-browser will continuously repeat the animation. This can be used, for example, to show the change of weather over the last day.

Slightly more interactive versions of this type of map are those to be played by media players, for instance those in *Quicktime* format. Plug-ins to the WWW-browser define the interaction options, which are often limited to simple pause, backward and forward play. Such animations do not use any specific WWW-environment parameters and have equal functionality in the desktop-environment. The WWW also allows for the fully interactive presentation of 3D models. The Virtual Reality Markup Language (VRML), is used for this, for instance. It stores a true 3D model of the objects, not just a series of 3D views.

Summary

Maps are the most efficient and effective means to inform us about spatial information. They 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 in and overview of the distribution of particular phenomena. An additional characteristic of particular on-screen maps is that they are often interactive and have a link to a database, and as such allow for more complicated queries.

Maps are the result of the visualization process. Their design is guided by "How do I say what to whom and is it effective?" Executing this sentence will inform the map maker about the characteristics of the data to be mapped, as well as the purpose of the map. This is necessary to find the proper symbology. The purpose could be to present the data to a wide audience or to explore the data to obtain better understanding. Cartographers have all kind of tools available to create appropriate visualizations. These tools consist of functions, rules and habits, together called the cartographic grammar.

This chapter discusses some characteristic mapping problems from the perspective of "How to map ..." First, the problem is described followed by a brief discussion of the potential solution based on cartographic rules and guidelines. The need to follow these rules and guidelines is illustrated by some maps that have been wrongly designed but are commonly found. The problems dealt with are "How to map qualitative data"—think of, for instance, soil or geological maps; "How to map quantitative data"—such as census data; "How to map the terrain"—dealing with relief, and informing about three-dimensional mapping options; "How to map time series"—such as urban growth presented in animations. Animations are well suited to display spatial change.

The map design will not only depend on the nature of the data to be mapped or the intended audience but also on the output medium. Traditionally, maps were produced on paper, and many still are. Currently, most maps are presented on screen, for a quick view, for an internal presentation or for presentation on the WWW. Each output medium has its own specific design criteria. All maps should have an appealing design and , next to the map image, have accessible a title, informing the user about the topic visualized.

Questions

1. Suppose one has two maps, one at scale $1 : 10,000$, and another at scale $1 : 1,000,000$. Which of the two maps can be called a large-scale map, and which a small-scale map?

2. Describe the difference between a topographic map and a thematic map.

3. Describe in one sentence, or in one question, the main problem of the cartographic visualization process.

4. Explain the content of Figure 6.8 in terms of that of Figure 3.1.
5. Which four main types of thematic data can be distinguished on the basis of their measurement scales?
6. Which are the six visual variables that allow to distinguish cartographic symbols from each other?

7. Describe a number of ways in which a three-dimensional terrain can be represented on a flat map display.

8. On page 383, we discussed three techniques for mapping changes over time. We already discussed the issue of change detection, and illustrated it in Figure 2.24. What technique was used there? Elaborate on how appropriate the two alternative techniques would have been in that example.

9. Describe different techniques of cartographic output from the user's perspective.

10. Explain the difference between static maps and dynamic maps.

Chapter 7

Data quality and metadata

7.1 Basic concepts and definitions

The purpose of any GIS application is to provide information to support planning and management. As this information is intended to reduce uncertainty in decision-making, any errors and uncertainties in spatial databases and GIS output products may have practical, financial and even legal implications for the user. For these reasons, those involved in the acquisition and processing of spatial data should be able to assess the *quality* of the base data and the derived information products.

Most spatial data are collected and held by individual, specialized organizations. Some 'base' data are generally the responsibility of the various governmental agencies, such as the National Mapping Agency, which has the mandate to collect topographic data for the entire country following pre-set standards. These organizations are, however, not the only sources of spatial data. Agencies such as geological surveys, energy supply companies, local government departments, and many others, all maintain spatial data for their own particular purposes. If this data is to be *shared* among different users, these users need to know not only what data exists, where and in what format it is held, but also whether the data meets their particular quality requirements. This 'data about data' is known as *metadata*.

This chapter has four purposes:

- to discuss the various aspects of spatial data quality,
- to explain how location accuracy can be measured and assessed,
- to introduce the concept of error propagation in GIS operations, and
- to explain the concept and purpose of metadata.

7.1.1 Data quality

The International Standards Organization (ISO) considers quality to be "the totality of characteristics of a product that bear on its ability to satisfy a stated and implied need" (Godwin, 1999). The extent to which errors and other shortcomings of a data set affect decision making depends on the *purpose* for which the data is to be used. For this reason, quality is often define as 'fitness for use'.

Traditionally, errors in paper maps are considered in terms of

1. *attribute* errors in the classification or labelling of features, and
2. errors in the *location*, or *height* of features, known as the *positional error*.

In addition to these two aspects, the International Cartographic Association's Commission on Spatial Data Quality, along with many national groups, has identified *lineage* (the history of the data set), *temporal accuracy*, *completeness* and *logical consistency* as essential aspects of spatial data quality.

In GIS, this wider view of quality is important for several reasons.

1. Even when source data, such as official topographic maps, have been subject to stringent quality control, errors are introduced when these data are input to GIS.
2. Unlike a conventional map, which is essentially a single product, a GIS database normally contains data from different sources of varying quality.
3. Unlike topographic or cadastral databases, natural resource databases contain data that are inherently uncertain and therefore not suited to conventional quality control procedures.
4. Most GIS analysis operations will themselves introduce errors.

7.1.2 Error

In day-to-day usage, the word *error* is used to convey that something is wrong. When applied to spatial data, error generally concerns mistakes or variation in the measurement of position and elevation, in the measurement of quantitative attributes and in the labelling or classification of features. Some degree of error is present in every spatial data set. It is important, however, to make a distinction between gross errors (blunders or mistakes), which ought to be detected and removed before the data is used, and the variation caused by unavoidable measurement and classification errors.

In the context of GIS, it is also useful to distinguish between errors in the *source data* and *processing errors* resulting from spatial analysis and modelling operations carried out by the system on the base data. The nature of positional errors that can arise during data collection and compilation, including those occurring during digital data capture, are generally well understood. A variety of tried and tested techniques is available to describe and evaluate these aspects of quality (see Section 7.2).

The acquisition of base data to a high standard of quality does not guarantee, however, that the results of further, complex processing can be treated with certainty. As the number of processing steps increases, it becomes difficult to predict the behaviour of this error propagation. With the advent of satellite remote sensing, GPS and GIS technology, 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. There is therefore a danger that uninformed GIS users introduce errors by wrongly applying geometric and other transformations to the spatial data held in their database.

7.1.3 Accuracy and precision

Measurement errors are generally described in terms of *accuracy*. The accuracy of a single measurement is

> "*the closeness of observations, computations or estimates to the true values or the values perceived to be true*" [48].

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. 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 in Section 7.2.4.

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. Prior to the availability of GPS, however, 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.

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 *accuracy tolerances*. In GIS, however, the numerical precision of computer processing and storage usually exceeds the accuracy of the data. This can give rise to so-called *spurious accuracy*, for example calculating area sizes to the nearest m^2 from coordinates obtained by digitizing a $1:50,000$ map.

7.1.4 Attribute accuracy

The 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 the 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 that can be used to evaluate the accuracy of the classification. An example is provided in Table 7.1, 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. For more details on attribute accuracy, the student is referred to Chapter 11 of *Principles of Remote Sensing* [30].

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 7.1: Example of a simple error matrix for assessing map attribute accuracy. The overall accuracy is $(62+18+12)/100 = 92\%$.

7.1.5 Temporal accuracy

In recent years, the amount of spatial data sets and archived remotely sensed data has increased enormously. 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*.

This 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 changes from time to time.

7.1.6 Lineage

Lineage describes the history of a data set. In the case of published maps, some lineage information may be provided in the form of a note on the data sources and procedures used in the compilation (for example, 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."* [15]

All of these aspects affect other aspects of quality, such as 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'.

7.1.7 Completeness

Data *completeness* is generally understood in terms of *omission errors*. The completeness of a map is a function of the cartographic and other procedures used in its compilation. The Spatial Data Transfer Standard (SDTS), and similar standards relating to spatial data quality, therefore includes information on classification criteria, definitions and mapping rules (for example, in generalization) in the statement of completeness.

Spatial data management systems—GIS, DBMS—accommodate some forms of incompleteness, and these forms come in two flavours. The first is a situation in which we are simply *lacking data*, for instance, because we have failed to obtain a measurement for some location. We have seen in previous chapters that operations of spatial inter- and extrapolation still allow us to come up with values in which we can have some faith.

The second type is of a slightly more general nature, and may be referred to as *attribute incompleteness*. It derives from the simple fact that we cannot know everything all of the time, and sometimes have to accept not knowing them. As this situation is so common, database systems allow to administer unknown attribute values as being *null*-valued. Subsequent queries on such (incomplete) data sets take appropriate action and treat the null values 'correctly'. Refer to Chapter 3 for details.

A form of incompleteness that is detrimental is positional incompleteness: knowing (measurement) values, but not, or only partly, knowing to what position they refer. Such data are essentially useless, as neither GIS nor DBMS systems accommodate them well.

7.1.8 Logical consistency

Completeness is closely linked to *logical consistency*, which deals with "the logical rules for spatial data and describes the compatibility of a datum with other data in a data set" [31]. Obviously, attribute data are also involved in a consistency question.

In practice, logical consistency is assessed by a combination of completeness testing and checking of topological structure as described in Section 2.2.4.

As previously discussed under the heading of database design, setting up a GIS and/or DBMS for accepting data involves a design of the data store. Part of that design is a definition of the data structures that will hold the data, *accompanied* by a number of rules of data consistency. These rules are dictated by the specific application, and deal with value ranges, and allowed combinations of values. Clearly, they can relate to both spatial and attribute data or arbitrary combinations of them. Important is that the rules are defined before any data is entered in the system as this allows the system to guard over data consistency from the beginning.

A few examples of logical consistency rules for a municipality cadastre application with a history subsystem are the following

- The municipality's territory is completely partitioned by mutually non-overlapping parcels and street segments. (A spatial consistency rule.)
- Any date stored in the system is a valid date that falls between January 1, 1900 and 'today'. (A temporal consistency rule.)
- The entrance date of an ownership title coincides with or falls within a month from the entrance date of the associated mortgage, if any. (A legal rule with temporal flavour.)

- Historic parcels do not mutually overlap in both valid time *and* spatial extent. (A spatio-temporal rule.)

Observe that these rules will typically vary from country to country—which is why we call them *application-specific*—but also that we can organize our system with data entry programs that will check all these rules automatically.

7.2 Measures of location error on maps

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 semantic (labelling) errors in their products, and, in addition, define quality in specifically cartographic terms, for example quality of linework, layout, and clarity of text.

All measurements made with surveying and photogrammetric instruments are subject to error. These include:

- human errors in measurement (e.g., reading errors),
- instrumental errors (e.g., due to misadjustment), and
- errors caused by natural variations in the quantity being measured.

7.2.1 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. The errors at each point can be plotted as error vectors, as is done in Figure 7.1 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.

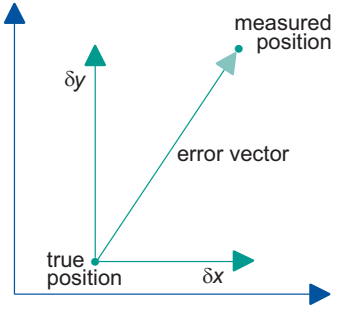


Figure 7.1: The positional error of a measurement can be expressed as a vector, which in turn can be viewed as the vector addition of its constituents in x- and y-direction, respectively δx and δy.

For each checkpoint, a vector can represent its location error. The 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 method of measuring. 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:

$$m_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \delta x_i^2} \text{ and } m_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \delta y_i^2},$$

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 indeed the length of the average (root squared) vector.

7.2.2 Accuracy tolerances

The *RMSE* can be used to assess the likelihood or 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 a normal (or Gaussian) distribution of a one-dimensional variable, 68.26% of the observed values lie within one standard deviation distance of the mean value. In the case of two-dimensional variables, like coordinates, the probability distribution takes the form of a bell-shaped surface (Figure 7.2). The three standard probabilities associated with this distribution are:

- 50% at $1.1774\ m_x$ (known as *circular error probable*, CEP);
- 63.21% at $1.412\ m_x$ (known as *root mean square error*, RMSE);

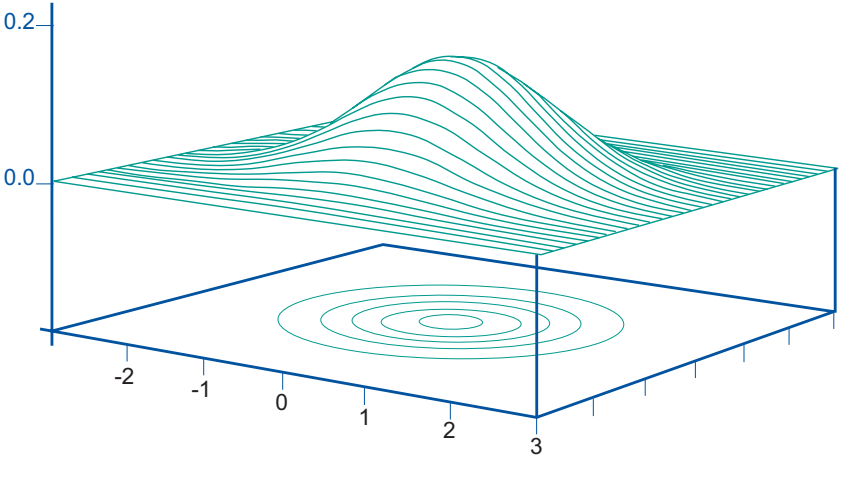


Figure 7.2: Probability of a normally distributed, two-dimensional variable (also known as a normal, bivariate distribution).

- 90% at $2.146\ m_x$ (known as *circular map accuracy standard*, 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.

7.2.3 The epsilon band

As a line is composed of an infinite number of points, confidence limits can be described by a so-called epsilon (ε) or Perkal band at a fixed distance on either side of the line (Figure 7.3). 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 7.4).

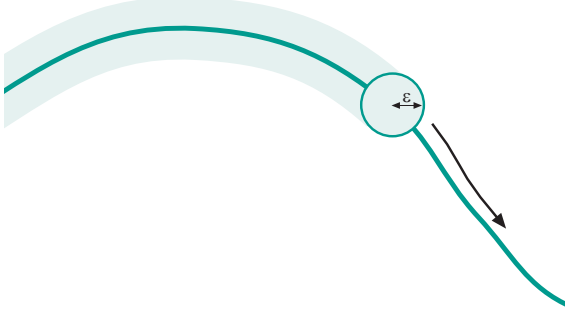


Figure 7.3: The ε- or Perkal band is formed by rolling an imaginary circle of a given radius along a line.

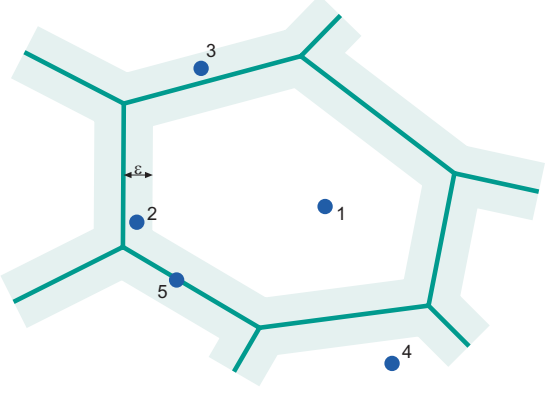


Figure 7.4: The ε-band may be used to assess the likelihood that a point falls within a particular polygon. Source: [50].

7.2.4 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" [10, 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.

Particularly in 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 cartographic accuracy, such as RMSE, may be virtually insignificant in comparison to the uncertainty inherent in, for example, vegetation and soil boundaries.

In conventional applications of the error matrix to assess the quality of nominal (categorical) coverages, such as land use, individual samples are considered in terms of Boolean set theory. The Boolean membership function is binary, i.e., an element is either member of the set (membership is `true`) or it is not member of the set (membership is `false`). Such a membership notion is well-suited to the description of spatial features such as land parcels where 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."

Since its original development by Zadeh [64], there has been considerable discussion of fuzzy, or continuous, set theory as an approach for handling im-

precise spatial data. In GIS, fuzzy set theory appears to have two particular benefits:

- the ability to handle logical modelling (map overlay) operations on inexact data, and
- 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 7.5). The membership function of the Boolean set of Figure 7.5(a) can be defined as MF^B follows:

$$\mathrm{MF}^B(x) = \begin{cases} 1 & \text{if } b_1 \leq x \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$

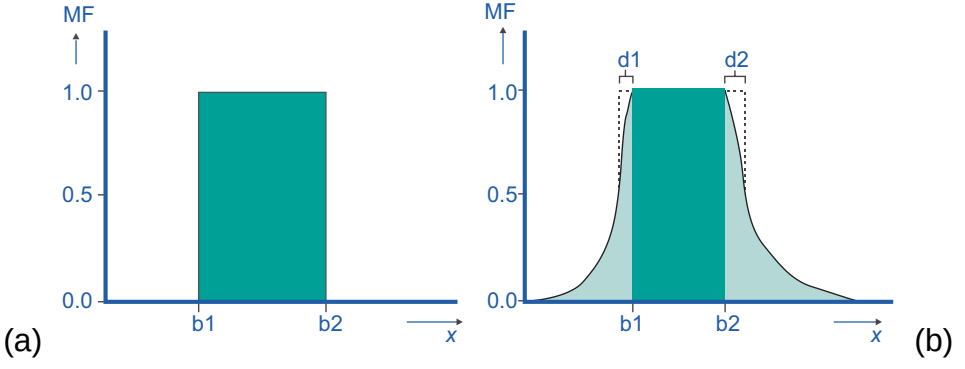


Figure 7.5: (a) Crisp (Boolean) and (b) uncertain (fuzzy) membership functions MF. After Heuvelink [25]

The crisp and uncertain set membership functions of Figure 7.5 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).

The continuous membership function of Figure 7.5(b), in contrast to function MF^B above, can be defined as a function MF^C, following Heuvelink in [25]:

$$\mathrm{MF}^C(x) = \begin{cases} \frac{1}{1+\left(\frac{x-b_1}{d_1}\right)^2} & \text{if } x < b_1 \\ 1 & \text{if } b_1 \leq x \leq b_2 \\ \frac{1}{1+\left(\frac{x-b_2}{d_2}\right)^2} & \text{if } x > b_2 \end{cases}$$

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 $b_1 - \frac{d_1}{2}$ and $b_2 + \frac{d_2}{2}$, 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," as such natural language expressions can be more faithfully represented by appropriately chosen membership functions.

7.3 Error propagation in spatial data processing

7.3.1 How errors propagate

In the previous section, we discussed a number of sources of error that may be present in source data. When these data are manipulated and analysed in a GIS, these various errors may affect the outcome of spatial data manipulations. The errors are said to *propagate* through the manipulations. In addition, further errors may be introduced during the various processing steps (see Figure 7.6).

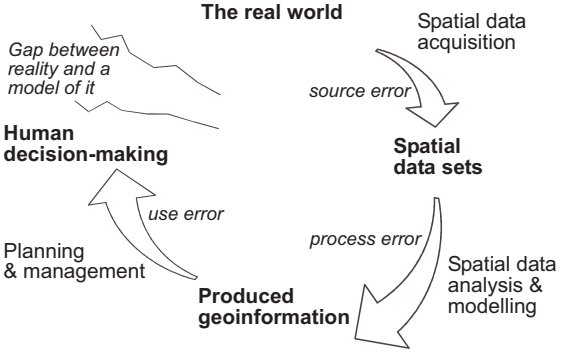


Figure 7.6: Error propagation in spatial data handling

For example, a land use planning agency may be 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 at two rainfall gauges.

The reader is invited to consider what sort of errors are likely to occur in this analysis.

One of the most commonly applied operations in geographic information systems is analysis by overlaying two or more spatial data layers. As discussed above, each such layer will contain errors, due to both inherent inaccuracies in the source data and errors arising from some form of computer processing, for example, rasterization. During the process of spatial overlay, all the 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. For example, errors in the results of overlay using the logical operator AND are not the same as those created using the OR operator.

7.3.2 Error propagation analysis

Two main approaches can be employed to assess 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.

Because "the ultimate arbiter of cartographic error is the real world, not a mathematical formulation" [14], there is much to recommend the use of testing procedures for accuracy assessment.

Models of error and error propagation

Modelling of error propagation has been defined by Veregin [62] as: "the application of formal mathematical models that describe the mechanisms whereby errors in source data layers are modified by particular data transformation operations." Thus, we would like to know how errors in the source data behave under manipulations that we subject them to in a GIS. If we somehow know 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 the results.

It is important to distinguish *models of error* from *models of error propagation* in GIS. Various perspectives, motives and approaches to 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.

Initially, the complexity of spatial data led to the development of mathematical models describing only the propagation of attribute error [25, 62]. More recent research has addressed the spatial aspects of error propagation and the development of models incorporating both attribute and locational components [3, 33]. All these approaches have their origins in academic research and have strong theoretical bases in mathematics and statistics. Although such technical work may eventually serve as the basis for routine functions to handle error and uncertainty, it may be argued that it is not easily understood by many of those using GIS in practice.

For the purpose of our discussion, we may look at a simple, arbitrary geographic field as a function A such that $A(x, y)$ is the value of the field in locality with coordinates (x, y). This field A may represent any continuous field: ground water salinity, soil fertility, or elevation, for instance. Now, when we discuss *error*, there is difference between what the actual value *is*, and what we *believe it to be*. What we believe is what we store in the GIS. As a consequence, if the *actual*

field is A, and our believe is the field B, we can write

$$A(x, y) = B(x, y) + V(x, y),$$

where $V(x, y)$ is the error in our approximation B at the locality with coordinates (x, y). This will serve as a basis for further discussion below. Observe that all that we know—and therefore have stored in our database or GIS—is B; we neither know A nor V.

Now, when we apply some GIS operator g—usually an overlay operator—on a number of geographic fields $A_1, \ldots, A_n$, in the ideal case we obtain an error-free output O_{ideal}:

$$O_{\text{ideal}} = g(A_1, \ldots, A_n). \tag{7.1}$$

Note that O_{ideal} itself is a geographic field. We have, however, just observed that we do not know the A_i's, and consequently, we cannot compute O_{ideal}. What we can compute is O_{known} as

$$O_{\text{known}} = g(B_1, \ldots, B_n),$$

with the B_i being the approximations of the respective A_i. The field O_{known} will serve as our approximation of O_{ideal}.

We wrote above that we do not know the actual field A nor the error field V. In most cases, however, we are not completely in the dark about them. Obviously, for A we have the approximation B already, while also for the error field V we commonly know at least a few characteristics. For instance, we may know with 90% confidence that values for V fall inside a range $[c_1, c_2]$. Or, we may know that the error field V can be viewed as a stochastic field that behaves in each locality (x, y) as having a normal distribution with a mean value $\overline{V}(x, y)$

and a variance $\sigma^2(x, y)$. The variance of V is a commonly used measure for data quality: the higher it is, the more variable the errors will be. It is with knowledge of this type that error propagation models may forecast the error in the output.

Models of error propagation based on first-order Taylor methods

It turns out that, unless drastically simplifying assumptions are made about the input fields A_i and the GIS function g, purely analytical methods for computing error propagation involve too high computation costs. For this reason, approximation techniques are much more practical. We discuss one of the simplest of these approximation techniques.

A well-known result from analytic mathematics, put in simplified words here, is the Taylor series theorem. It states that a function $f(z)$, if it is differentiable in an environment around the value $z = a$, can be represented within that environment as

$$f(z) = f(a) + f'(a)(z-a) + \frac{f''(a)}{2!}(z-a)^2 + \frac{f'''(a)}{3!}(z-a)^3 + \cdots \tag{7.2}$$

Here, f' is the first, f'' the second derivative, and so on.

In this section, we use the above theorem for computing O_{ideal}, which we defined in Equation 7.1. Our purpose is not to find the O_{ideal} itself, but rather to find out what is the effect on the resulting errors.

In the *first-order Taylor method*, we deliberately make an approximation error, by ignoring all higher-order terms of the form $\frac{\cdots}{n!}(z-a)^n$ for $n \geq 2$, assuming that they are so small that they can be ignored. We apply the Taylor theorem with function g for placeholder f, and the vector of stored data sets $(B_1, \ldots, B_n)$ for placeholder a in Equation 7.2. As a consequence, we can write

$$O_{\text{ideal}} = g(B_1, \ldots, B_n) + \sum_{i=1}^{n} (A_i - B_i) g'_i(B_1, \ldots, B_n),$$

Under these simplified conditions, it can be shown that the mean value for O_{ideal}, viewed as a stochastic field, is $g(B_1, \ldots, B_n)$. In other words, we can use

the result of the g computation on the stored data sets as a sensible predictor for O_{ideal}.

It has also been shown, what the above assumptions mean for the variance of stochastic field O_{ideal}, denoted by τ^2. The formula that [25] derives is:

$$\tau^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} \sigma_i \sigma_j g'_i(B_1, \ldots, B_n) g'_j(B_1, \ldots, B_n),$$

where ρ_{ij} denotes the correlation between input data sets B_i and B_j and σ_i^2, as before, is the variance of input data set B_i.

The variance of O_{ideal} (under all mentioned assumptions) can be computed and depends on a number of factors: the correlations between input data sets, their inherent variances, as well as the steepness of the function g. It is especially this steepness that may cause our resulting error to be 'worse' or not.

7.4 Metadata and data sharing

Over the past 25 years, spatial data has been collected in digital form at increasing rate, and stored in various databases by the individual producers for their own use and for commercial purposes. These data sets are usually in miscellaneous types of store that are not well-known to many.

The rapid development of information technology—with GIS as an important special case—has led to an increased pressure on the people that are involved in analysing spatial data and in providing such data to support decision making processes. This prompted these data suppliers to start integrating already existing data sets to deliver their products faster. Processes of spatial data acquisition are rather costly and time consuming, so efficient production is of a high priority.

7.4.1 Data sharing and related problems

Geographic data exchange and sharing means the flow of digital data from one information system to the other. Advances in technology, data handling and data communication allow the users to think of the possibility of finding and accessing data that has been collected by different data providers. Their objective is to minimize the duplication of effort in spatial data collection and processing. Data sharing as a concept, however, has many inherent problems, such as

- the problem of locating data that are suitable for use,
- the problem of handling different data formats,
- other heterogeneity problems, such as differences in software (versions),
- institutional and economic problems, and finally
- communication problems.

Data distribution

Spatial data are collected and kept in a variety of formats by the producers themselves. What data exists, and where and in what format and quality the data is available is important knowledge for data sharing. These questions, however, are difficult to answer in the absence of a utility that can provide such information. Some base data are well known to be the responsibility of various governmental agencies, such as national mapping agencies. They have the mandate to collect topographic data for the entire country, following some standard. But they are not the only producers of spatial data.

Questions concerning quality and suitability for use require knowledge about the data sets and such knowledge usually is available only inside the organization. But if data has to be shared among different users, the above questions need to be addressed in an efficient way. This data about data is what is commonly referred to as 'metadata'.

Data standards

The phrase 'data standard' refers to an agreed upon way of representing data in a system in terms of content, type and format. Exchange of data between databases is difficult if they support different data standards or different query languages. The development of a common data architecture and the support for a single data exchange format, commonly known as *standard for data exchange* may provide a sound basis for data sharing. Examples of these standards are the Digital Geographic Information Exchange Standard (DIGEST), Topologically Integrated Geographic Encoding and Referencing (TIGER), Spatial Data Transfer Standard (SDTS).

The documentation of spatial data, i.e. the metadata, should be easy to read and understand by different discipline professionals. So, standards for metadata are also required.

These requirements do not necessarily impose changing the existing systems, but rather lead to the provision of additional tools and techniques to facilitate data sharing. A number of tools have been developed in the last two decades to harmonize various national standards with international standards. We devote a separate section (Section 7.4.2) to data standards below.

Heterogeneity

Heterogeneity means being different in kind, quality or character. Spatial data may exist in a variety of locations, are possibly managed by a variety of database systems, were collected for different purposes and by different methods, and are stored in different structures. This brings about all kinds of inconsistency among these data sets (heterogeneity) and creates many problems when data is shared.

Institutional and economic problems

These problems arise in the absence of policy concerning pricing, copyright, privacy, liability, conformity with standards, data quality, etc. Resolving these problems is essential to create the right environment for data sharing.

Communication problems

With advances in computer network communication and related technology, locating relevant information in a network of distributed information sources has become more important recently. The question is which communication technology is the best suitable for transfer of 'bulk'—i.e., huge amounts of—spatial data in a secure and reliable way. Efficient tools and communication protocols are necessary to provide search, browse and delivery mechanisms.

7.4.2 Spatial data transfer and its standards

The need to exchange data among different systems leads to the definition of standards for the transfer of spatial data. The purpose of these transfer standards is to move the contents of one GIS database to a different GIS database with a minimal loss of structure and information. Since the early 1980s, many efforts have been made to develop such standards on a local, national and international level. Today, we have operational transfer standards that support the dissemination of spatial data.

In a transfer process, data is always physically moved from one system to the other. A completely different approach to data sharing is interoperability of GIS. Here, GIS software accesses data on different systems (connected via a computer network) through standardized interfaces. Data does not need to be physically converted and transferred. The Open GIS Consortium is the leading organization that coordinates these activities. It is a consortium of major GIS and database software vendors, academia and users.

When we transfer data between systems, we might encounter various problems: transfer media are not compatible, physical and logical file formats could be different, or we might not have any information concerning the quality of the data set. Moreover, we need translators from and into every format that we might have to deal with. In the worst case, for n different systems we need $n(n-1)$ translators. The solution to these problems is to exchange spatial data in a standardized way thereby keeping as much as possible of the structure and relationships among the features in the data set. Using a standard reduces the number of required translators to $2n$, because we need one translator—from the GIS database to the standard, and back—for each system.

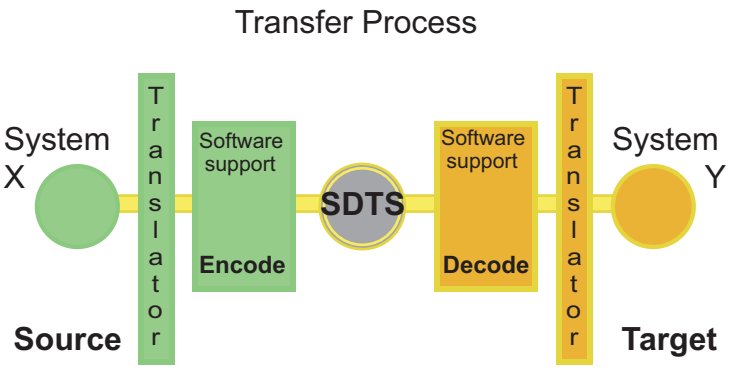


Figure 7.7: Spatial data transfer process

The spatial data transfer process

A data transfer standard defines a data model for the transfer as well as a translation mechanism from a GIS to the exchange format and into the exchange mechanism. In transferring data from system X to system Y, we need a translator that converts the contents of the database in system X into the model of the spatial data transfer standard (SDTS).[1] The components of the model are represented by modules that are converted into a computer readable format. A standard often used for this exchange format is the ISO 8211 Data Descriptive File.

On the receiving end, system Y needs a translator that converts the data from the transfer standard into the database. In the ideal case, after the transfer is completed, no further processing is needed. The data are ready to use (Figure 7.7).

[1]Here, we use SDTS as the abbreviation for the generic term spatial data transfer standard. This should not be confused with the FIPS 173 SDTS, the federal spatial data transfer standard of the United States of America.

Examples of spatial data transfer standards

Standards are accepted either as industry and *de facto* standards, or as authoritative national or international standards. Industry standards are frequently used standards that were introduced by a company or organization but which are not accepted as national or federal standards. Examples of such standards are the USGS DLG (digital line graph of the United States Geological Survey), or the DXF file formats (AutoCAD Format).

Authoritative standards are accepted as federal or national standards based on international ISO standards. The following Table 7.2 gives an overview of some transfer standards in different countries and organizations.

NATO	DIGEST (Digital Geographic Exchange Standard) of the Digital Geographic Information Working Group (DGIWG); members are Belgium, Canada, Denmark, France, Germany, Italy, The Netherlands, Norway, Spain, United Kingdom, USA
Europe	Geographic information – Data description – Transfer (CEN ENV 12658:1998)
United Kingdom	NTF (National Transfer Format), Electronic transfer of geographic information, UK, BS 7567, 1992
USA	SDTS (Spatial Data Transfer Standard), FIPS 173, 1992
France	Echanges de Données Informatisés dans le domaine de l'information GéOgraphique - EDIGéO (AFNOR, Z 13-150, 1992)

Table 7.2: Examples of spatial data transfer standards

7.4.3 Geographic information infrastructure and clearinghouses

The design of an infrastructure that facilitates the discovery of sources of geographic information is the focus of action in many countries. *Geographic information infrastructure* (GII), also referred to as Spatial Data Infrastructure (SDI), can be defined as a collection of institutional, economic and technical tools arranged in a way that improves the timely accessibility to required information. These tools should help to resolve the problems listed above.

A *formal data resource* is an integrated, comprehensive data source that makes data readily identifiable and easily accessible. Looking at the world today, a clearinghouse plays such a role of formal data resource.

A (spatial data) *clearinghouse* is a distributed network of spatial data producers, managers and users that are linked electronically together. It is a system of software and institutions that are to facilitate the discovery, evaluation, and downloading of digital spatial data and provides means to inventory, document and data sharing. The *clearinghouse* concept is a useful one in building a GII. The objective is to minimize unnecessary duplication of effort for data capture, and to maximize the benefit of geographic information sharing.

Data providers nowadays are fully aware of the importance of advertising and making available their metadata describing their databases, to facilitate the use of their products. This explains the current level of activity of building these clearinghouses.

How does a clearinghouse work?

A clearinghouse allows data providers to register their geographic data sets, the quality of these data and also the instructions for accessing them. Each data provider provides an electronic description of each spatial data set. In addition, the provider may also provide access to the spatial data set itself. The clearinghouse thus functions as a detailed catalogue service with support for links to spatial data and browsing capabilities. The data described in the clearinghouse may be located at the site of the data producers or at sites of designated data disseminators located elsewhere in the country. Obviously, computer networks facilities are the key factor to success.

7.4.4 Metadata concepts and functionality

The Information Age has, in the past decades, produced a new vocabulary of terms and concepts by which to describe data and information. Discussions on metadata focus on issues of adequate description, standardized format and ease of locating. Such issues must conform to international standards.

Metadata is defined as background information that describes the content, quality, condition and other appropriate characteristics of the data. Metadata is a simple mechanism to inform others of the existence of data sets, their purpose and scope. In essence, metadata answer *who, what, when, where, why*, and *how* questions about all facets of the data made available.

Metadata can be used internally by the data provider to monitor the status of data sets, and externally to advertise to potential users through a national clearinghouse. Metadata are important in the production of a digital spatial data clearinghouse, where potential users can search for the data they need.

Metadata play a variety of informative roles:

Availability: information needed to determine the data sets that exist for a geographic location,

Fitness for use: information needed to determine whether a data set meets a specified need,

Access: information needed to acquire an identified data set,

Transfer: information needed to process and use a data set,

Administration: information needed to document the status of existing data (data model, quality, completeness, temporal validity, *et cetera*) to define internal policy for update operations from different data sources.

The metadata should be flexible enough to describe a wide range of data types. Details of the metadata vary with the purpose of their use, so certain levels of abstraction are required.

Metadata standards

For metadata to be easily read and understood, standards create a common language for users and producers. Metadata standards provide appropriate and adequate information for the design of metadata.

Key developments in metadata standards are the ISO STANDARD 15046-15 METADATA, the Federal Geographic Data Committee's content standard for Digital Geospatial Metadata FGDC, CSDGM, the European organization responsible for standards CEN/TC 287 and others. Several studies have been conducted to show how data elements from one standard map into others.

A standard provides a common terminology and definitions for the documentation of spatial data. It establishes the names of data elements and groups of data elements to be used for these purposes, the definitions of these data elements and groups, and information about the values that can be assigned to the data elements. Information about terms that are mandatory, mandatory under certain conditions, or optional (provided at the discretion of the data provider) also are defined in the standard.

The choice of which metadata standard to use depends on the organization, the ease of use and the intended purpose.

Definitions of data elements

The FGDC standard specifies the structure and expected content of more than 220 items. These are intended to describe digital spatial data sets adequately for all purposes. They are grouped into seven categories:

Identification Information: basic information about the data set. Examples include the title, the geographic area covered, currentness, and rules for acquiring or using the data.

Data Quality Information: an assessment of the quality of the data set. Examples include the positional and attribute accuracy, completeness, consistency, the sources of information, and methods used to produce the data. Recommendations on information to be reported and tasks to be performed are in the Spatial Data Transfer Standard (Federal Information Processing Standard 173).

Spatial Data organization Information: the mechanism used to represent spatial information in the data set. Examples include the method used to represent spatial positions directly (such as raster or vector) and indirectly (such as street addresses or county codes) and the number of spatial objects in the data set.

Spatial Reference Information: description of the spatial reference frame for, and means of, encoding coordinates in the data set. Examples include the name of and parameters for map projections or grid coordinate systems, horizontal and vertical datums, and the coordinate system resolution.

Entity and Attribute Information: information about the content of the data set, including the entity types and their attributes and the domains from which

attribute values may be assigned. Examples include the names and definitions of features, attributes, and attribute values.

Distribution Information: information about how the data set can be acquired. For instance, a contact address of the distributor, available formats, information about how to obtain data sets on-line or on physical media (such as cartridge tape or CD-ROM), and fees for the data.

Metadata Reference Information: information on the currentness of the metadata information and the responsible party.

The standard has sections that specify contact information for organizations or individuals that developed or distribute the data set, temporal information for time periods covered by the data set, and citation information for the data set and information sources from which the data set was derived.

Metadata management and update

Just like ordinary data, metadata has to be kept up-to-date. The main concerns in metadata management include what to represent, how to represent it, how to capture and how to use it; and all these depend on the purpose of the metadata:

For internal (data provider) use, we will refer to 'local metadata', which contains the detailed information about data sets stored on local hardware and managed by the data provider. For external use, we refer to 'global metadata', which contains a short description of the data sets (an abstraction of the local metadata) as advertised in the clearinghouse to allow users to find relevant data efficiently.

Data providers should register their data holding with the clearinghouse. Whenever changes occur in their data, each data provider reports the changes to the clearinghouse authority. Updating the global metadata is the responsibility of the clearinghouse.

7.4.5 Structure of metadata

Metadata can be structured or unstructured. Unstructured metadata consist of free-form textual descriptions of data and processes. Structured metadata consist mainly of relationship definitions among the data elements. Structured metadata is important as it can be indexed and searched, moreover, its is much easier to exchange with others.

All proposed standards for metadata provide well defined items that can be used to judge fitness for use, to order and to use the data sets.

Summary

The essential function of a GIS is to produce information with the aim of reducing uncertainty in management and decision making. Reliable information implies that the base data meets defined standards of *quality*. Quality is therefore defined as '*fitness for use.*' A quality statement for a spatial data set should include information on:

- *Lineage* (the history of the data set),
- *Positional accuracy*, for example, the RMSE of check measurements,
- *Attribute accuracy*, such as an error matrix based on field checking of maps made from remotely sensed sources,
- *Completeness* of the data set, and
- *Logical consistency* of the data set.

Quality information is an important component of *metadata*, that is 'data about data'. Metadata is increasingly important as digital data are *shared* among different agencies and users. Metadata include basic information about:

- *What data exist* (the content and coverage of a data set),
- The *quality* of the data,
- The *format* of the data, and
- Details about how to obtain the data, its cost, and restrictions on its use.

Questions

1. List three source errors and three processing errors. (See page 402.)

2. The following data show the surveyed coordinates of twelve points and their 'true' values as obtained from check measurements of a higher order of accuracy. Assume the values given below are in metres.

Easting (x)	Measured	Northing (y)	Measured
21215	21216	18785	18785
21235	21233	18787	18786
21233	21230	18765	18763
21265	21266	18783	18782
21291	21291	18770	18769
21217	21219	18746	18746
21254	21253	18753	18752
21287	21286	18748	18747
21224	21226	18727	18725
21235	21237	18717	18718
21254	21253	18726	18724
21276	21276	18717	18715

   (a) Calculate the error at each point.

   (b) Check if there is a systematic error.

   (c) Calculate m_x, m_y and the total RMSE.

   (d) Plot the positions of the points at a scale of $1 : 1000$.

(e) Plot the error vectors at a scale of $1 : 100$.

3. In which situations are spatial data transfer standards relevant for a GIS application? When does one need to know about these standards?

4. Try to find—on the Internet—a spatial data clearinghouse, and identify what sort of data can be obtained through it. In a second stage, reverse the search, and first identify a spatial data set that you would like to obtain, then try find it. The more specific your requirements, the more difficult obviously the search. Relax the requirements if necessary, but pay attention to which relaxations pay off.

Bibliography

[1] Edwin A. Abbott. *Flatland—A romance of many dimensions*. Penguin Group, New York, N.Y., 1984.

[2] A. Agumya and G. J. Hunter. Determining fitness for use of geographic information. *ITC Journal*, 1997(2):109–113, 1997.

[3] 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. 426

[4] Stan Aronoff. *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa, Canada, 1989. 147, 158, 278

[5] A. S. Belward and C. R. Valenzuela, editors. *Remote Sensing and Geographical Information Systems for Resource Management in Developing Countries*. Kluwer Academic, Dordrecht, The Netherlands, 1991.

[6] J. Bertin. *Sémiology Graphique*. Mouton, Den Haag, The Netherlands, 1967. 365, 370

[7] Wietske 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. 124, 127

[8] C. Board. Report of the working group on cartographic definitions. *Cartographic Journal*, 29(1):65–69, 1990. 352

[9] Graeme F. Bonham-Carter. *Geographic information systems for geoscientists : Modeling with GIS*, volume 13 of *Computer methods in the geosciences*. Pergamon, Kidlington, U.K., 1994. 306

[10] P. A. Burrough. Natural objects with indeterminate boundaries. In P. A. Burrough and A. U. Frank, editors, *Geographic objects with indeterminate boundaries*, pages 3–28. Taylor and Francis, London, U.K., 1996. 419

[11] Peter A. Burrough. *Principles of Geographical Information Systems for Land Resources Assessment*. Monographs on Soil and Resources Survey. Clarendon Press, Oxford, U.K., 1986.

[12] Peter A. Burrough and R. McDonnell. *Principles of Geographical Information Systems*. Oxford University Press, Oxford, U.K., 1998.

[13] W. Cartwright, M. Peterson, and G. Gartner, editors. *Multimedia Cartography*. Springer, Berlin, Germany, 1999.

[14] N. R. Chrisman. Errors in categorical maps: testing versus simulation. In *Proceedings AutoCarto*, pages 521–529, 1989. 425

[15] 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. 408

[16] Chris J. Date. *An Introduction to Database Systems*. Addison-Wesley, Reading, Ma, seventh edition, 2000. 186

[17] Borden D. Dent. *Cartography: Thematic Map Design*. WCB/McGraw-Hill, Boston, Ma, fifth edition, 1999.

[18] D. DiBiase. Visualization in earth sciences. *Earth and Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences*, 59(2):13–18, 1990. 362

[19] M. Ehlers and S. Amer. Geoinformatics: An integrated approach to acquisition, processing and production of geo-data. In *Proceedings EGIS'91, Brussels*, pages 306–312, 1991. 144

[20] Ramez Elmasri and Shamkant B. Navathe. *Fundamentals of Database Systems*. Benjamin/Cummings, Redwood City, Ca, second edition, 1994. 186

[21] L. Godwin. Establishing quality principles. *GIM*, 13(8):6–9, 1999.

[22] Michael F. Goodchild. Geographical information science. *Int. J. Geographical Information Systems*, 6(1):31–45, 1992. 144

[23] Richard Groot. Meeting educational requirements in geomatics. *ITC Journal*, 1989(1):1–4, 1989. 144

[24] H. M. Hearnshaw and D. J. Unwin, editors. *Visualization in Geographical Information Systems*. John Wiley & Sons, London, U.K., 1994.

[25] Gerard B. M. Heuvelink. *Error propagation in quantitative spatial modelling—Applications in Geographical Information Systems*. Nederlandse Geografische Studies. Koninklijk Aardrijkskundig Genootschap, Utrecht, 1993. 420, 421, 426, 430

[26] S. W. Houlding. *3D Geoscience Modeling: Computer Techniques for Geological Characterization*. Springer-Verlag, Berlin, Germany, 1994.

[27] Jonathan Iliffe. *Datums and Map Projections for Remote Sensing, GIS and Surveying*. Whittles Publishing, CRC Press, 2000. 227

[28] ILWIS Department. *ILWIS 2.1 for Windows—User's Guide*. ITC, Enschede, The Netherlands, 1997.

[29] Intergovernmental Committee on Surveying and Mapping (ICSM). Get in step with the geocentric datum. Technical report, Office of the Surveyor General, Land Information New Zealand (LINZ), 1999. 219

[30] Lucas L. F. Janssen and Gerrit C. Huurneman, editors. *Principles of Remote Sensing*, volume 2 of *ITC Educational Textbook Series*. International Institute for Aerospace Survey and Earth Sciences, Enschede, The Netherlands, second edition, 2001. 18, 35, 65, 118, 196, 237, 405

[31] 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. 410

[32] Wolfgang Kainz, Max Egenhofer, and I. Greasley. Modeling spatial relations and operations with partially ordered sets. *Int. J. Geographical Information Systems*, 7(3):215–229, 1993.

[33] H. T. Kiiveri. Assessing, representing and transmitting positional accuracy in maps. *International Journal of Geographical Information Systems*, 11(1):33–52, 1997. 426

[34] Donald E. Knuth. *The TEXbook*. Addison-Wesley, Reading, Ma, 1984. 21

[35] Menno-Jan Kraak. Exploratory cartography, maps as tools for discovery. *ITC Journal*, 1998(1):46–54, 1998.

[36] Menno-Jan Kraak and Allan Brown, editors. *Web cartography, developments and prospects*. Taylor & Francis, London, U.K., 2000. 350, 357, 363, 365, 387, 391

[37] Menno-Jan Kraak and F. J. Ormeling. *Cartography: Visualization of Spatial Data*. Addison-Wesley Longman, London, U.K., 1996. 354, 361

[38] Land Information New Zealand (LINZ). Strategic business plan. Technical report, Land Information New Zealand (LINZ), October 1999. 219

[39] Gail Langran. *Time in Geographic Information Systems*. Technical Issues in Geographic Information Systems. Taylor & Francis, London, U.K., 1992. 128

[40] Robert Laurini and Derek Thompson. *Fundamentals of Spatial Information Systems*, volume 37 of *The APIC Series*. Academic Press, London, U.K., 1992. 144

[41] Mary-Claire van Leunen. *A Handbook for Scholars*. Oxford University Press, New York, NY, revised edition, 1992. 21

[42] 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. 361

[43] A. M. MacEachren and D. R. F. Taylor, editors. *Visualization in Modern Cartography*. Pergamon Press, London, U.K., 1994.

[44] Bruce McCormick, Tomas A. DeFanti, and Maxine D. Brown (eds.). Visualization in scientific computing. *ACM SIGGRAPH Computer Graphics—Special issue*, 21(6), 1987. 362

[45] A. M. J. Meijerink, J. A. M. de Brouwer, C. M. Mannaerts, and C. R. Valenzuela. *Introduction to the Use of Geographic Information Systems for Practical Hydrology*, volume 23 of *ITC Publication*. ITC, Enschede, The Netherlands, 1994.

[46] Martien Molenaar. *An Introduction to the Theory of Spatial Object Modelling*. Taylor & Francis, London, U.K., 1998.

[47] 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. 363

[48] National Mapping Division, U. S. Geological Survey. Spatial data transfer standard. Technical report, U. S. Department of the Interior, 1990. 403

[49] Office of the Surveyor General of Land Information New Zealand. A proposal for geodetic datum development. Technical Report OSG TR2.1, Land Information New Zealand (LINZ), 1998. 219

[50] 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. 418

[51] Donna J. Peuquet and D. F. Marble, editors. *Introductory Readings in Geographic Information Systems*. Taylor & Francis, London, U.K., 1990.

[52] Reinhard Pflug and John W. Harbaugh, editors. *Three-dimensional Computer Graphics in Modeling Geologic Structures and Simulating Geologic Processes*, volume 41 of *Lecture Notes in Earth Sciences*. Springer-Verlag, Berlin, Germany, 1992.

[53] Franco P. Preparata and Michael I. Shamos. *Computational Geometry—An Introduction*. Springer-Verlag, New York, NY, 1985. 93

[54] Jonathan Raper, editor. *Three dimensional Applications in Geographic Information Systems*. Taylor & Francis, London, U.K., 1989.

[55] A. H. Robinson, J. L. Morrison, P. C. Muehrcke, A. J. Kimerling, and S. C. Guptill. *Elements of Cartography*. John Wiley & Sons, New York, N.Y., sixth edition, 1995. 361

[56] Hanan Samet. *Applications of Spatial Data Structures*. Addison-Wesley, Reading, Ma, 1990.

[57] Hanan Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, Ma, 1990. 156

[58] J. Star and J. Estes. *Geographic Information Systems, An Introduction*. Prentice Hall, Englewood Cliffs, NJ, 1990.

[59] William Strunk Jr. and E. B. White. *The Elements of Style*. MacMillan Publishing Company, New York, NY, third edition, 1979. 21

[60] C. Dana Tomlin. *Geographic Information Systems and Cartographic Modeling*. Prentice Hall, Englewood Cliffs, NJ, 1990.

[61] A. K. Turner, editor. *Three-dimensional Modeling with Geoscientific Information Systems*. Kluwer Academic, Dordrecht, The Netherlands, 1992.

[62] 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. 426

[63] Michael F. Worboys. *GIS: A Computing Perspective*. Taylor & Francis, London, U.K., 1995.

[64] Lofty A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965. 419

[65] Michael Zeiler. *Modeling our World—The ESRI Guide to Geodatabase Design*. ESRI Press, Redlands, Ca, 1999.

Appendix A

Internet sites

General GIS sites

- Institut Géographique National, France (in French)
- Geoplaza Netherlands (in Dutch)
- United States Geological Survey (USGS)
- GISLinx.com, list of GIS sites
- University of Iowa, Center for Global and Regional Environmental Research
- L'Institut Lorrain de Génie Urbain (ILGU), Nancy, France
- Oddens' Bookmarks, Faculty of Geographical Sciences, Utrecht University, The Netherlands

Spatial data sources

- Digital Chart of the World, at Pennsylvania State University, U.S.A.
- Pennsylvania State University Libraries, Maps Library
- U.S.A. Federal Geographic Data Committee Clearinghouse
- United States Geological Survey (USGS) Geospatial Data Clearinghouse
- ESRI's Data Repository

Spatial reference systems and frames

- Geometric Aspects of Mapping; Division of Cartography, ITC.
- Active GPS Reference System for the Netherlands (AGRS.NL)
- SAtelliten POsitionierung System (SAPOS)
- Deutsches Geodätisches Forschungsinstitut (DGFI), Bayerische Akademie der Wissenschaften. Geodetic Reference System 1980 (GRS80).
- Office of the Surveyor General of Land Information New Zealand (LINZ). A Proposal for Geodetic Datum Development. OSG TR2.1, 1998
- Ordnance Survey of Great Britain. A Guide to Coordinate Systems in Great Britain, 2000.

Glossary

Abbreviations & Foreign words

2D two-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).

$2\frac{1}{2}$D 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 $2\frac{1}{2}$D system can at most associate only one elevation. A TIN structure, for instance, is a typical $2\frac{1}{2}$D structure, as it only determines single elevation values for single locations.

3D three-dimensional. Typically applied to (aspects of) GIS applications that view their phenomena in a three-dimensional space, where coordinates are triplets (x, y, z).

ADSL Asymmetric Digital Subscriber Lines. A new technology of data transmission used to deliver high-rate digital data over existing ordinary phone-lines. ADSL facilitates the simultaneous use of normal telephone services, ISDN, and high speed data transmission, e.g., video.

a posteriori in retrospect; when looking back.

a priori (assumed known) beforehand.

ArcInfo A GIS software package developed in the 1980s and 90s at ESRI. As

the name indicates ('Arc'), historically more vector-based than raster-based.

ASCII American Standard Code for Information Interchange; an encoding of text characters into integer values represented as bytes. So-called 'plain text' files usually are encoded in ASCII.

AVHRR Advanced Very High Resolution Radiometer; a broad-band scanner, sensing in the visible, near-infrared, and thermal infrared portions of the electromagnetic spectrum, carried on NOAA's Polar Orbiting Environmental Satellites (POES).

bps bits per second. The unit in which data transmission rates are measured. Eight bits constitute a byte, which is used to represent a single character in a text document. The usual unit is now Mbps: million bits per second. A data rate of 1Mbps allows to transmit about 40 pages of plain text per second.

DBMS Database Management System.

Digital Elevation Model (DEM) Special case of a DTM. A DEM stores terrain elevation (surface height) by means of a raster. The word 'elevation' refers to a height expressed with respect to a specific reference.

Digital Terrain Model (DTM) Term indicating a digital description of the terrain relief. A DTM can be stored in different manners (contour lines, TIN, raster) and may also contain semantic, relief-related information (breaklines, saddlepoints).

dpi dots per inch; the unit of scanner (or printer) resolution, expressed as how many pixels can be read (printed) per inch.

e.g. for example, ; (*exempli gratia*).

ESRI Environmental Systems Research Institute, Inc. The American company that sells ArcInfo and ArcView.

GII Geographic Information Infrastructure; sometimes also known as Spatial Data Infrastructure (SDI).

GIS Geographic Information System.

i.e. that is, ; meaning, ; (*id est*).

ILWIS Integrated Land and Water Information System. A GIS software package developed in the 1980s and 1990s at ITC. Historically more raster-based than vector-based.

in situ on the spot; in the terrain.

ISO International Standards Organization.

ITRF International Terrestrial Reference Frame.

ITRS International Terrestrial Reference System.

NOAA National Oceanic Atmospheric Administration; an institute falling under the U.S. Department of Commerce, aiming, amongst others through satellite imagery, at monitoring the Earth's environment.

RMSE Root Mean Square Error.

SDI Spatial Data Infrastructure; see GII.

SQL Structured Query Language; the query language implemented in all relational database management systems.

SRF Spatial Reference Frame.

SRS Spatial Reference System.

SST Sea Surface Temperature; as used in examples of Chapter 1.

TIN Triangulated Irregular Network.

viz. namely, ; (*videlicet*).

WS Wind Speed; as used in examples of Chapter 1.

WWW World-wide Web. In a broad sense, the global internet with all the information and services that can be found there.

Terms

accuracy closeness (i.e., degree of match) of measurements, observations, computations or estimates to the (perceived to be) true values.

attribute the name of a column in a database table; it should suggest what the values in that column stand for. These values are known as *attribute values*.

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*.

buffer area surrounding a selected set of features. May be defined in terms of a fixed distance, or by a more complicated relationship that the features may have on their surroundings.

cartography the whole of scientific, technological and artistic activities directed to the conception, production, dissemination and use of map displays.

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.

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*.

contour map map in which contour lines are used to represent terrain elevation.

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*.

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.

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.

dynamic map (also: cartographic animation); map with changing contents, and/or changing ways of representation of these contents, whether triggered by the user or not.

epoch (precise) date and time. Used to register at what moment a measurement took place, in this book, the moment at which the measurements took place for fixing ('freezing') the positions of the fundamental polyhedron of a spatial reference frame.

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 commonest Euclidean space in GIS use.

Evapotranspiration (sometimes erroneously written as evapotranspo*r*ation); the process by which surface water, soils, and plants release water vapour to the atmosphere through evaporation (surface water, solis) and transpiration (plants).

exploratory cartography interactive cartographic visualization of not well-understood spatial data by an individual to stimulate visual thinking and to create insight in and overview of the spatial data.

feature collective noun to indicate either a point, polyline or polygon vector object, when the distinction is not important.

geographic dimension Spatial phenomena exist in space and time. The geographic dimension is the space factor in this existence, and determines *where* the phenomenon is present.

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 Information derived from spatial data. Strictly speaking, information is derived by humans using mental processes, so geographic information too is made of mental 'matter' only. Day-to-day use of the term, however, allows us to exchange it with 'spatial data'.

geographic information system A software package that accommodates the capture, analysis, manipulation and presentation of georeferenced data. It is a generic tool applicable to many different types of use (GIS applications).

geographic phenomenon Any man-made or natural phenomenon (that we are interested in).

geographic space Space in which locations are defined relative to the Earth's surface. The usual space that GIS applications work with.

georeferenced Data is georeferenced when coordinates from a geographic space have been associated with it. The georeference (spatial reference) tells us where the object represented by the data is (or was or will be). (As such, an abbreviation of 'geographically referenced'.).

geospatial data Data that includes positions in geographic space. In this book, usually abbreviated to 'spatial data'.

geovisualization making spatial data 'visible' by means of maps generated through interactive and dynamic software tools.

GIS application Software specifically developed to support the study of geographic pheneomena in some application domain in a specific project. A spatial data set as stored in a GIS, together with functions on the data. Serves a well-defined purpose, making use of GIS functionality. Distinguished from the software—the GIS package, the database package—that can be applied generically.

granularity The level of detail with which something is represented.

grid A regularly spaced set of points with associated (field) values, defined as the intersections of perpendicular gridlines.

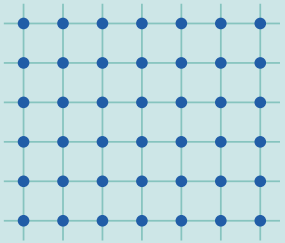


Figure A.1: A grid is a collection of regularly spaced points. The associated values with each point are not illustrated.

In contrast to a *raster*, the associated values represent *point* values, not cell values. This subtlety is often—and can often be—glossed over, especially when point distances are small relative to the variation in the represented phenomenon. By default a grid is two-dimensional, but we can think of three-dimensional grids.

integer any 'whole' number in the set $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$; computers cannot represent arbitrarily large numbers, and some maximum (and minimum) integer is usually indicated.

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.

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.

line A computer representation of a geographic object that is perceived as a one-dimensional, i.e., curvilinear entity. The line determines two end nodes plus a, possibly empty, list of internal points, known as vertices. Other words for 'line' are polyline (emphasising the multiple linear segments), arc or edge.

lineage the recorded history of a spatial data set or spatial data product, including where relevant details about the production process that allow to assess characteristics of data quality.

man-made phenomenon An object, occurrence or event that was created by humans. This is a difficult to define and large population of entities: anything that can be georeferenced and originates from man can be a 'man-made phenomenon'.

map originally, a reduced and simplified representation of a chosen set of geographic phenomena in a planar display; nowadays, still that, but

new technical possibilities allow to look at more generic map notions, for instance, including virtual reality and time awareness.

map generalisation the meaningful reduction of map content to accommodate scale decrease.

map projection the functional mapping of a curved horizontal reference surface onto a flat 2D plane, using mathematical equations.

map scale ratio between a distance on the map and the corresponding distance in reality.

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.

natural phenomenon An object, occurrence or event that originated naturally. This is a difficult to define and large population: see also 'man-made phenomenon' as a contrast.

nominal data data values that serve to identify or name something, but that do not allow arithmetic computations; sometimes also called categorical data.

oblate ellipsoid the solid (i.e., a three-dimensional object) produced by rotating an ellipse (i.e., a two-dimensional object) about its minor axis. It is also known as *spheroid*, because it resembles a sphere flattened (squashed) at the poles.

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.

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.

polyhedron a solid bounded by planar facets, i.e., a three-dimensional feature of which the sides are flat surfaces. The *fundamental polyhedron* of the ITRF is a mesh of foundation stations around the globe that are used to define the ITRS.

presentation cartography cartographic visualization of spatial data for presentation to a group of users (public visual communication).

raster A regularly spaced set of cells with associated (field) values.

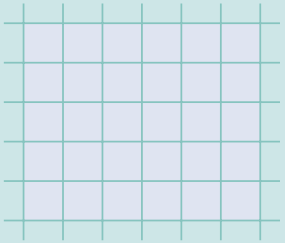


Figure A.2: A raster is a regularly spaced set of cells. The associated values with each cell are not illustrated.

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

vald for all locations within the cell. This subtlety is often—and can often be—glossed over, especially when the cell size is small relative to the variation in the represented phenomenon. By default a raster is two-dimensional, but we can think of three-dimensional rasters as well.

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.

simplex A primitive spatial feature as recognized in topology. A 0-simplex is a point, 1-simplex an arc, a 2-simplex an area and a 3-simplex a body. See *simplicial complex*.

simplicial complex A combination, i.e. spatial arrangement, of a number of simplices, possibly of different dimension.

solid a true three-dimensional object.

spatial data In the precise sense, spatial data is any data with which position is associated. In this book, we use the phrase mostly as 'geospatial data', meaning that geographic position data is part of it.

spatial data layer a collection of data items that belong together, and that can be spatially interpreted. A raster is a spatial data layer, and so are a collection of polygons, a collection of polylines, or a collection of point features. Principles of correct data organization dictate that the raster's cells (or the polygons, polylines or points) represent phenomena of the same kind.

spatial database A database that, amongst others, stores 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 reference frame A physical realisation of a spatial reference system, consisting of real point objects (ground stations) with their coordinates in the used SRS. In fact, next to the coordinates for each object also of the object's motion in time, due to tectonic plate movement, is recorded.

spatial reference system A 3D reference coordinate system with well-defined origin and orientation of the coordinate axes. A mathematical system.

spatial relationship A mathematically defined relationship between two simplicial complices (objects), usually defining whether they are disjoint, meet, overlap *et cetera*. Spatial relationships are the object of study in topology.

sphere the solid (i.e., a three-dimensional object) produced by rotating a circle.

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.

string any sequence of characters chosen from the alphabet plus a set of other characters like interpunction symbols ('?', '!', ';' *et cetera*) and numbers. When typed to a computer, a string is usually surrounded by a pair of double quotes.

temporal dimension Spatial phenomena exist in space and time. The temporal dimension is the time factor in this existence, and determines *when* the phenomenon is present.

tessellation (also known as 'tiling'); a partition of space into mutually disjoint cells that together form the complete study area. A raster is a regular tessellation example, meaning that its constituent cells have the same shape and size. In irregular tessellations, the cells differ in shape and/or in size.

thematic map a map in which the distribution, quality and/or quantity of a phenomenon (or the relationship among several phenomena) is presented on a topographic base.

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.

topographic map a map that gives a general, realistic and complete, but simplified representation to scale of the terrain (roads, rivers, buildings and settlements, vegetation, relief, geographical names, *et cetera*.).

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 The mathematical study of the properties of space(s) and invariance characteristics under space transformations.

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. The underlying principle is that the locations constitute the corner points of a collection of triangles that is a spatial partition of the study area. The field value for an arbitrary location is interpolated from the values of the corner points of the triangle inside which that location falls.

triangulation A complete partition of the study space into mutually non-overlapping triangles, usually on the basis of georeferenced measurements.

tuple a record or row in a database table; it will have several attribute values. Pronounce as 'tapl'.

visual variable (also: graphic variable); an elementary way in which graphic symbols are distinguished from each other. Commonly, the follow-

ing six visual variables are recognized: *size*, *(lightness) value*, *texture*, *colour*, *orientation* and *shape*.

Index