

Principles of Geographic Information Systems

An introductory textbook

Editors

Otto Huisman and Rolf A. de By

Cover illustration:

Paul Klee (1879–1940), *Chosen Site* (1927)

Pen-drawing and water-colour on paper. Original size: 57.8 × 40.5 cm.

Private collection, Munich

© Paul Klee, Chosen Site, 2001 c/o Beeldrecht Amstelveen

Cover page design: Wim Feringa

All rights reserved. No part of this book may be reproduced or translated in any form, by print, photoprint, microfilm, microfiche or any other means without written permission from the publisher.

Published by:

The International Institute for Geo-Information Science and Earth Observation
(ITC),

Hengelosestraat 99,

P.O. Box 6,

7500 AA Enschede, The Netherlands

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Principles of Geographic Information Systems

Otto Huisman, Rolf A. de By (eds.)

(ITC Educational Textbook Series; 1)

[previous](#)

[next](#)

[back](#)

[exit](#)

[contents](#)

[index](#)

[glossary](#)

[web links](#)

[bibliography](#)

[about](#)

Fourth edition

ISBN 978-90-6164-269-5 ITC, Enschede, The Netherlands

ISSN 1567-5777 ITC Educational Textbook Series

© 2009 by ITC, Enschede, The Netherlands

Contents

1 A gentle introduction to GIS	25
1.1 The nature of GIS	26
1.1.1 Some fundamental observations	29
1.1.2 Defining GIS	32
1.1.3 GISystems, GIScience and GIS applications	43
1.1.4 Spatial data and geoinformation	45
1.2 The real world and representations of it	48
1.2.1 Models and modelling	49
1.2.2 Maps	51
1.2.3 Databases	53
1.2.4 Spatial databases and spatial analysis	55
1.3 Structure of this book	57
2 Geographic information and Spatial data types	62
2.1 Models and representations of the real world	63
2.2 Geographic phenomena	66
2.2.1 Defining geographic phenomena	67

2.2.2	Types of geographic phenomena	69
2.2.3	Geographic fields	72
2.2.4	Geographic objects	77
2.2.5	Boundaries	81
2.3	Computer representations of geographic information	82
2.3.1	Regular tessellations	85
2.3.2	Irregular tessellations	88
2.3.3	Vector representations	91
2.3.4	Topology and spatial relationships	101
2.3.5	Scale and resolution	113
2.3.6	Representations of geographic fields	114
2.3.7	Representation of geographic objects	119
2.4	Organizing and managing spatial data	124
2.5	The temporal dimension	126
3	Data management and processing systems	135
3.1	Hardware and software trends	137
3.2	Geographic information systems	140
3.2.1	GIS software	142
3.2.2	GIS architecture and functionality	144
3.2.3	Spatial Data Infrastructure (SDI)	146
3.3	Stages of spatial data handling	148
3.3.1	Spatial data capture and preparation	149
3.3.2	Spatial data storage and maintenance	151
3.3.3	Spatial query and analysis	155
3.3.4	Spatial data presentation	157
3.4	Database management systems	158

3.4.1	Reasons for using a DBMS	160
3.4.2	Alternatives for data management	163
3.4.3	The relational data model	164
3.4.4	Querying a relational database	171
3.5	GIS and spatial databases	178
3.5.1	Linking GIS and DBMS	179
3.5.2	Spatial database functionality	182
4	Spatial referencing and positioning	189
4.1	Spatial referencing	191
4.1.1	Reference surfaces for mapping	192
4.1.2	Coordinate systems	206
4.1.3	Map projections	217
4.1.4	Coordinate transformations	227
4.2	Satellite-based positioning	236
4.2.1	Absolute positioning	238
4.2.2	Errors in absolute positioning	246
4.2.3	Relative positioning	254
4.2.4	Network positioning	256
4.2.5	Code versus phase measurements	257
4.2.6	Positioning technology	258
5	Data entry and preparation	270
5.1	Spatial data input	271
5.1.1	Direct spatial data capture	272
5.1.2	Indirect spatial data capture	274
5.1.3	Obtaining spatial data elsewhere	280

5.2	Data quality	284
5.2.1	Accuracy and precision	285
5.2.2	Positional accuracy	287
5.2.3	Attribute accuracy	298
5.2.4	Temporal accuracy	300
5.2.5	Lineage	301
5.2.6	Completeness	302
5.2.7	Logical consistency	303
5.3	Data preparation	304
5.3.1	Data checks and repairs	305
5.3.2	Combining data from multiple sources	312
5.4	Point data transformation	320
5.4.1	Interpolating discrete data	323
5.4.2	Interpolating continuous data	325
6	Spatial data analysis	342
6.1	Classification of analytical GIS capabilities	344
6.2	Retrieval, classification and measurement	349
6.2.1	Measurement	350
6.2.2	Spatial selection queries	355
6.2.3	Classification	368
6.3	Overlay functions	376
6.3.1	Vector overlay operators	377
6.3.2	Raster overlay operators	381
6.3.3	Overlays using a decision table	390
6.4	Neighbourhood functions	392
6.4.1	Proximity computations	395

6.4.2	Computation of diffusion	400
6.4.3	Flow computation	403
6.4.4	Raster based surface analysis	405
6.5	Network analysis	415
6.6	GIS and application models	424
6.7	Error propagation in spatial data processing	429
6.7.1	How errors propagate	430
6.7.2	Quantifying error propagation	434
7	Data visualization	440
7.1	GIS and maps	441
7.2	The visualization process	452
7.3	Visualization strategies: present or explore?	456
7.4	The cartographic toolbox	463
7.4.1	What kind of data do I have?	464
7.4.2	How can I map my data?	466
7.5	How to map ...?	470
7.5.1	How to map qualitative data	471
7.5.2	How to map quantitative data	473
7.5.3	How to map the terrain elevation	477
7.5.4	How to map time series	481
7.6	Map cosmetics	485
7.7	Map dissemination	490
Glossary		504
A Internet sites		536

List of Figures

1.1	The El Niño event of 1997 compared with a normal year 1998	30
1.2	Schema of an SST measuring buoy	35
1.3	The array of measuring buoys	36
1.4	Just four measuring buoys	60
2.1	Three views of objects of study in GIS	64
2.2	Elevation as a geographic field	71
2.3	Geological units as a discrete field	74
2.4	Geological faults as geographic objects	79
2.5	Three regular tessellation types	85
2.6	A grid and a raster illustrated	86
2.7	An example region quadtree	89
2.8	Input data for a TIN construction	92
2.9	Two triangulations from the same input data	93
2.10	An example line representation	97
2.11	An example area representation	98
2.12	Polygons in a boundary model	100

2.13	Example topological transformation	102
2.14	Simplices and a simplicial complex	105
2.15	Spatial relationships between two regions	108
2.16	The five rules of topological consistency in 2D space	110
2.17	Raster representation of a continuous field	115
2.18	Vector representation of a continuous field	117
2.19	Image classification of an agricultural area	120
2.20	Image classification of an urban area	121
2.21	A straight line and its raster representation	122
2.22	Geographic objects and their vector representation	123
2.23	Overlaying different rasters	124
2.24	Producing a raster overlay layer	125
2.25	Change detection from radar imagery	130
3.1	Functional components of a GIS	145
3.2	Example relational database	165
3.3	Example foreign key attribute	170
3.4	The two unary query operators	175
3.5	The binary query operator	176
3.6	A combined query	177
3.7	Raster data and associated database table	180
3.8	Vector data and associated database table	181
3.9	Geometry data stored in spatial database	183
4.1	Two reference surfaces for approximating the Earth	192
4.2	The Geoid	193
4.3	Levelling network	195
4.4	An oblate ellipse	196

4.5	Regionally best fitting ellipsoid	198
4.6	Dutch triangulation network	201
4.7	The ITRS and ITRF	202
4.8	Height above the geocentric ellipsoid and above the Geoid	205
4.9	2D geographic coordinate system	207
4.10	3D geographic coordinate system	209
4.11	3D geocentric coordinate system	211
4.12	2D cartesian coordinate system	212
4.13	Coordinate system of the Netherlands	214
4.14	2D polar coordinate system	215
4.15	Projecting geographic into cartesian coordinates	218
4.16	Classes of map projections	221
4.17	Three secant projection classes	221
4.18	A transverse and an oblique projection	222
4.19	Mercator projection	224
4.20	Cylindrical equal-area projection	225
4.21	Equidistant cylindrical projection	226
4.22	Changing map projection	230
4.23	Changing projection combined with a datum transformation	232
4.24	Determining pseudorange and position	239
4.25	Satellite positioning	242
4.26	Positioning satellites in view	250
4.27	Geometric dilution of precision	253
4.28	GPS satellite constellation	260
5.1	The phases of the vectorization process	278
5.2	Good/bad accuracy against good/bad precision	286

5.3	The positional error of a measurement	289
5.4	A normally distributed random variable	291
5.5	Normal bivariate distribution	292
5.6	The ε - or Perkal band	294
5.7	Point-in-polygon test with the ε -band	294
5.8	Crisp and uncertain membership functions	297
5.9	Successive clean-up operations for vector data	307
5.10	The integration of two vector data sets may lead to slivers	313
5.11	Multi-scale and multi-representation systems compared	316
5.12	Multiple adjacent data sets can be matched and merged	317
5.13	Interpolating quantitative and qualitative measurements	321
5.14	Generation of Thiessen polygons for qualitative data	324
5.15	Various global trend surfaces	327
5.16	Interpolation by triangulation	331
5.17	The principle of moving window averaging	332
5.18	Inverse distance weighting as an averaging technique	334
6.1	Minimal bounding boxes	352
6.2	Interactive feature selection	357
6.3	Spatial selection through attribute conditions	358
6.4	Further spatial selection through attribute conditions	359
6.5	Spatial selection using containment	363
6.6	Spatial selection using intersection	364
6.7	Spatial selection using adjacency	365
6.8	Spatial selection using the distance function	366
6.9	Two classifications of average household income per ward	369
6.10	Example discrete classification	372

6.11	Two automatic classification techniques	375
6.12	The polygon intersect overlay operator	377
6.13	The residential areas of Ilala District	378
6.14	Two more polygon overlay operators	380
6.15	Examples of arithmetic map algebra expressions	383
6.16	Logical expressions in map algebra	386
6.17	Complex logical expressions in map algebra	387
6.18	Examples of conditional raster expressions	389
6.19	The use of a decision table in raster overlay	391
6.20	Buffer zone generation	396
6.21	Thiessen polygon construction from a Delaunay triangulation .	399
6.22	Diffusion computations on a raster	401
6.23	Flow computations on a raster	404
6.24	Moving window rasters for filtering	410
6.25	Slope angle defined	411
6.26	Slope angle and slope aspect defined	413
6.27	Part of a network with associated turning costs at a node	417
6.28	Ordered and unordered optimal path finding	419
6.29	Network allocation on a pupil/school assignment problem	421
6.30	Tracing functions on a network	422
6.31	Error propagation in spatial data handling	430
7.1	Maps and location	442
7.2	Maps and characteristics	443
7.3	Maps and time	444
7.4	Comparing aerial photograph and map	445
7.5	Topographic map of Overijssel	449

7.6	Thematic maps	450
7.7	Dimensions of spatial data	451
7.8	Cartographic visualization process	452
7.9	Visual thinking and communication	458
7.10	The cartographic communication process	461
7.11	Bertin's six visual variables	467
7.12	Qualitative data map	471
7.13	Two wrongly designed qualitative maps	472
7.14	Mapping absolute quantitative data	473
7.15	Two wrongly designed quantitative maps	474
7.16	Mapping relative quantitative data	475
7.17	Bad relative quantitative data maps	476
7.18	Visualization of the terrain	479
7.19	Quantitative data in 3D visualization	480
7.20	Mapping change	484
7.21	The map and its information	487
7.22	Text in the map	488
7.23	Visual hierarchy	489
7.24	Classification of maps on the WWW	491

List of Tables

1.1	Average sea surface temperatures in December 1997	40
1.2	Database table of daily buoy measurements	54
3.1	Commonly used unit prefixes	138
3.2	Spatial data input methods and devices used	150
3.3	Raster and vector representations compared	152
3.4	Spatial data presentation	157
3.5	Three relation schemas	167
4.1	Three global ellipsoids	199
4.2	Transformation of Cartesian coordinates	234
4.3	Transformation from the Potsdam datum	235
4.4	Magnitude of errors in absolute satellite-based positioning	252
5.1	A simple error matrix	299
5.2	Clean-up operations for vector data	306
6.1	Example continuous classification table	370

6.2	Common causes of error in spatial data handling	433
7.1	Data nature and measurement scales	465
7.2	Measurement scales linked to visual variables	469

Preface

This book was originally designed for a three-week lecturing module on the principles of Geographic Information Systems (GIS), 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 supports the study of natural and man-made phenomena with an explicit location in space. To this end, the GIS allows data entry, data manipulation, and production of interpretable output that may provide new insights about the phenomena.

There are many uses for GIS technology, including 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 educated in one of these fields; the intention of the book is to lay the foundation for the reader to also become proficient in the use of GIS technology.

With so many different fields of application, it is impossible to single out the specific techniques of GIS usage for all the fields in a single book. Rather, the

book focuses on a number of common and important topics that any expert GIS user should be aware of. GIS is a continuously evolving scientific discipline, and for this reason ITC students should be provided with a broad foundation of relevant concepts, techniques and technology.

The book is also meant to define a common understanding and terminology for follow-up modules, which the student may elect later in her/his respective programme. The textbook does not stand independently, but was developed in conjunction with the textbook on *Principles of Remote Sensing*.

Structure of this book

The chapters of the book have been arranged in a semi-classical set-up. Chapters 1 to 3 provide a general introduction to the field, discussing various interesting geographic phenomena (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). Spatial referencing and positioning (including map projections and GPS) is dealt with in Chapter 4.

Chapters 5 to 7 subsequently focus on the *process* of using a GIS environment. We discuss how spatial data can be obtained, entered and prepared for use (Chapter 5), how data can be manipulated to improve our understanding of the phenomena that they represent (Chapter 6), and how the results of such manipulations can be visualized (Chapter 7). Special attention throughout these chapters is devoted to the specific characteristics of *geospatial* data.

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, and an index. The book is also made available as an electronic PDF document which can be browsed but not printed.

Acknowledgements

The book has a significant history, dating from the 1999 curriculum. This is a heavily revised, in parts completely rewritten, version of that first edition.

Major contributors to the current content of this book include Rolf de By, Richard Knippers, Michael Weir, Yola Georgiadou, Menno-Jan Kraak, Cees van Westen and Yuxian Sun. Authors of the original text include Martin Ellis, Wolfgang Kainz, Mostafa Radwan and Ed Sides.

Other contributors which deserve a great deal of credit for their management, assistance and/or advisory roles in the production of previous editions of this book include Erica Weijer, Marion van Rinsum, Ineke ten Dam, Kees Bronsveld, Rob Lemmens, Connie Blok, Allan Brown, Corné van Elzakker, Lucas Janssen, Barend Köbben, Bart Krol, and Jan Hendrikse.

Facilitated by technical advice from Wim Feringa, many illustrations in the book were 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.

For this fourth edition, a number of colleagues provided valuable comments and prepared materials that helped to substantially improve the existing text. They include Richard Knippers, Connie Blok, Ellen-Wien Augustijn, Rob Lemmens, Wim Bakker, Ivana Ivanova, Nicholas Hamm, Jelger Kooistra, Chris Hecker and Karl Grabmaier. Wim Feringa deserves special thanks for providing new and

redrawn figures for this edition, as well as his work on the updated cover design.

The editors would also like to acknowledge the pleasant collaboration with Klaus Tempfli, the editor of *Principles of Remote Sensing* and Coco Rulinda for L^AT_EX issues.

Technical account

This book was written using Leslie Lamport's L^AT_EX generic typesetting system, which uses Donald Knuth's T_EX 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 L^AT_EX sources we generated the book in PDF format, using the PDFL^AT_EX macro package, supported by various add-on packages, the most important being Sebastian Rahtz' hyperref.

Preface to the fourth edition

This fourth edition of the GIS book is an update of the previous edition, with some reshuffling of the book's content and some minor changes to the layout (marginal editorial disagreements notwithstanding). Care has been taken to provide updated material, improve readability and browse-ability of the text, and achieve greater integration through cross-referencing.

Significant changes include a rewritten section on spatial referencing by Richard Knippers, restructuring of material in chapters 3, 4 and 5, and a range of edits for improved continuity throughout the chapters. A keyword-in-the-margin layout was adopted to aid in browse-ability of the main text.

It must be stressed that the design of this book remains that of a textbook on 'principles'. A much bigger overhaul would have been required for another format, and this was considered undesirable for its purpose, and infeasible in the time allowed, also because of dependencies with already developed teaching materials such as exercises and overhead slides.

This textbook continues to be used at ITC in all educational programmes, as well as in other programmes around the globe that are developed in collaboration with ITC. A Korean translation of both textbooks has already been published, and other translation projects are under way. People with an interest in such an undertaking are invited to contact the editors.

A book such as this will never be perfect, and the field of GIScience has not yet reached the type of maturity where debates over definitions and descriptions are no longer needed. The Editors welcome any comments and criticisms, in a

continued effort to improve the materials.

Otto Huisman and Rolf A. de By, Enschede, July 2009

Chapter 1

A gentle introduction to GIS

[previous](#)[next](#)[back](#)[exit](#)[contents](#)[index](#)[glossary](#)[web links](#)[bibliography](#)[about](#)

1.1 The nature of GIS

The purpose of this chapter is to set the scene for the remainder of this book by providing a general overview of some of the terms, concepts and ideas which will be covered in greater detail in later sections.

The acronym GIS stands for *geographic information system*. As the name suggests, a GIS is a tool for working with geographic information. Section 1.1.2 provides a more formal definition, and later sections will look in more detail at some of the key functions that set GIS apart from other kinds of information systems. GIS have rapidly developed since the late 1970's in terms of both technical and processing capabilities, and today are widely used all over the world for a wide range of purposes. Let us begin by looking at some of these:

Geographic information
system

- An *urban planner* might want to assess the extent of 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 *these* particular suburbs are growing and others are not;
- 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 long-term threats to those populations;
- A *natural hazard analyst* might like to identify the high-risk areas of annual monsoon-related flooding by investigating rainfall patterns and terrain characteristics;

- A *geological engineer* might want to identify the best localities for constructing buildings in an earthquake-prone area by looking at rock formation characteristics;
- A *mining engineer* could be interested in determining which prospective copper mines should be selected 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 telecommunications 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 need to preserve species diversity in the area;
- A *hydrological engineer* might want to study a number of water quality parameters of different sites in a freshwater lake to improve understanding of the current distribution of *Typha* reed beds, and why it differs from that of a decade ago.

In the examples presented above, all the professionals work with positional data – also called *spatial data*. Spatial data refers to *where* things are, or perhaps, where they were or will be. To be more precise, these professionals deal with questions related to *geographic space*, which we define as having positional data relative to the Earth's surface.

Spatial data

Positional data of a non-geographic nature also exists. Examples include the location of the appendix in the human body, or the location of headlights on a car. These examples involve positional information, but it makes no sense to use the Earth's surface as a reference for these applications. For the purposes of this book we are only interested in *geographic* data. To illustrate these issues further, the following section provides an example of the application of GIS to the study of global weather patterns.

Geographic data

1.1.1 Some fundamental observations

Our world is dynamic. Many aspects of our daily lives and our environment are constantly changing, and not always for the better. Some of these changes appear to have natural causes (e.g. volcanic eruptions, meteorite impacts), while others are the result of human modification of the environment (e.g. land use changes or land reclamation from the sea, a favourite pastime of the Dutch). There are also a large number of global changes for which the cause remains unclear: these include global warming, the El Niño/La Niña events, or at smaller scales, landslides and soil erosion. 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 fully understand the changes.

Dynamics and change

For background information on El Niño, please refer to Figure 1.1. This Figure presents information related to a study area (the equatorial Pacific Ocean), with positional data taking a prominent role. Although quite a complex phenomenon, we will use the study of El Niño as an example application of GIS in the remainder of this chapter.

In order to understand what is going on in our world, we study the processes or *phenomena* that bring about geographic change. In many cases, we want to broaden or deepen our understanding to help us make decisions, so that we can take the best course of action. For instance, if we understand El Niño better, and can forecast that another event may take place in the year 2012, 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.

Geographic phenomena

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', and manifests itself usually around Christmas. There exists also another—less pronounced—pattern of *colder* temperatures, that is known as La Niña ('little girl') which occurs less frequently than El Niño. The most recent occurrence of El Niño started in September 2006 and lasted until early 2007. From June 2007 on, data indicated a weak La Niña event, strengthening in early 2008. 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, an 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, while the bottom figures are labelled with 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.

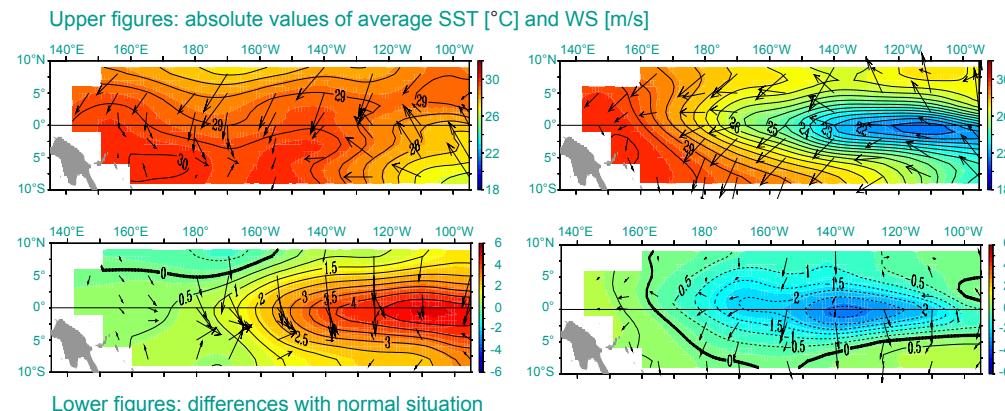


Figure 1.1: The El Niño event of 1997 compared with a more normal year 1998. The top figures indicate average Sea Surface Temperature (SST, in colour) and average Wind Speed (WS, in arrows) for the month of December. The bottom figures illustrate the anomalies (differences from a normal situation) in 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).

The fundamental problem that we face in many uses of GIS is that of understanding phenomena that have a spatial or *geographic dimension*, as well as a *temporal dimension*. We are facing 'spatio-temporal' problems. This means that our object of study has different characteristics for different locations (the geographic dimension) and also that these characteristics change over time (the temporal dimension). The El Niño event is a good example of such a phenomenon, because sea surface temperatures differ between locations, and sea surface temperatures change from one week to the next.

Spatial and temporal dimensions

1.1.2 Defining GIS

The previous section illustrated the use of GIS in a range of settings to operate on data that represent geographic phenomena. This provides us with a functional definition (after Aronoff [3]):

A GIS is a computer-based system that provides the following four sets of capabilities to handle georeferenced data:

1. *Data capture and preparation*
2. *Data management*, including storage and maintenance
3. *Data manipulation and analysis*
4. *Data presentation*

This implies that a GIS user can expect support from the system to enter (georeferenced) data, to analyse it in various ways, and to produce presentations (including maps and other types) from the data. This would include support for various kinds of coordinate systems and transformations between them, options for analysis of the georeferenced data, and obviously a large degree of freedom of choice in the way this information is presented (such as colour scheme, symbol set, and medium used).

For examples of each of these capabilities, let us take a closer look at the El Niño example. Many professionals closely study this phenomenon, most notably meteorologists and oceanographers. They prepare all sorts of products, such as the

maps of Figure 1.1, in order to improve their understanding. To do so, they need to obtain data about the phenomenon, which, as shown above, includes measurements about sea water temperature and wind speed from many locations. This data must be stored and processed to enable it to be analysed, and allow the results from the analysis to be interpreted. The way this data is presented could play an important role in its interpretation.

We have listed these capabilities above in the most natural order in which they take place. But this is only a sketch of an ideal situation, and it is often the case that data analysis suggests that we need more data about the problem. Data presentation may also lead to follow-up questions for which we need to do more analysis, and for which we may need more data, or perhaps better data. Consequently, several of the steps may be repeated a number of times before we are happy with the results. We look into these steps in more detail below, in the context of the El Niño example.

Data capture and preparation

In the El Niño case, data capture refers to the collection of sea water temperatures and wind speed measurements. This is achieved by placing buoys with measuring equipment at various places in the ocean. Each buoy measures a number of things: wind speed and direction; air temperature and humidity; and sea water temperature at the surface and at various depths down to 500 metres. For the sake of our example we will focus 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. For monitoring purposes, some 70 buoys were deployed at strategic places within 10° latitude of the Equator, between the Galápagos Islands and Papua 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 by high seas and bad weather or by the buoys becoming entangled in long-line fishing nets.¹

All the data that a buoy obtains through its thermometers and other sensors, as well as the buoy's geographic position are transmitted by satellite communication daily. Later in this book, and also in the textbook on *Principles of Remote Sensing* [53], many other ways of acquiring geographic data will be discussed.

¹As Figure 1.3 shows, there happen to be three types of buoy, but their differences are not directly relevant to our example, so we will ignore them here.

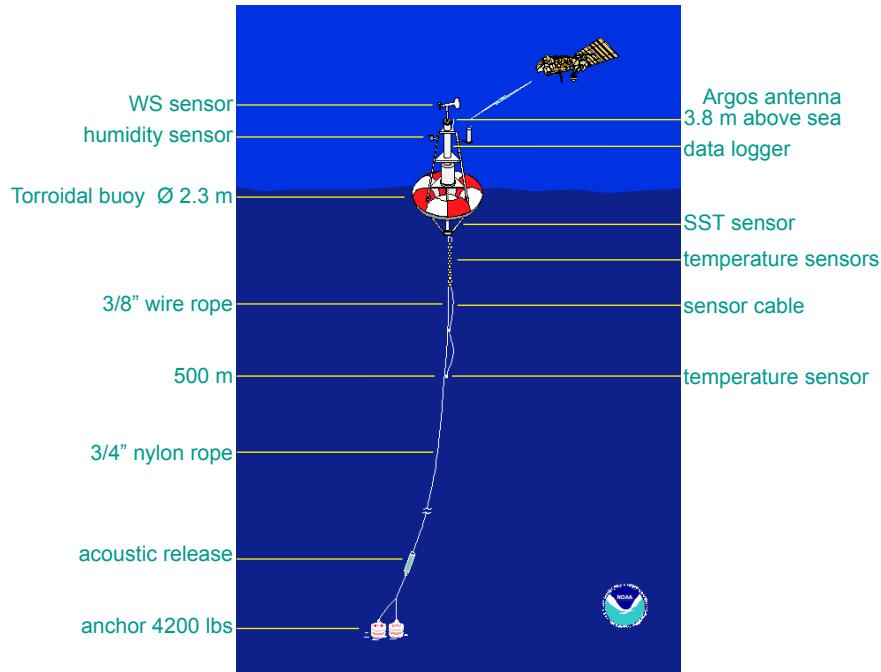


Figure 1.2: Schematic overview of an ATLAS type buoy for monitoring sea water temperatures in the El Niño project

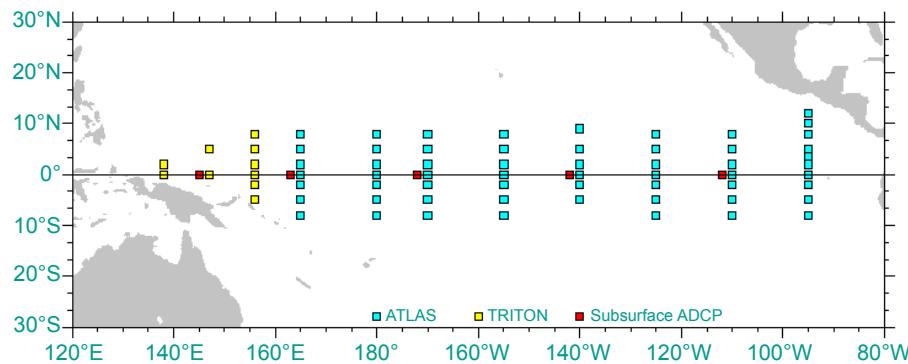


Figure 1.3: The array of positions of sea surface temperature and wind speed measuring buoys in the equatorial Pacific Ocean

Data management

For our example application, data management refers to the storage and maintenance of the data transmitted by the buoys via satellite communication. This phase requires a decision to be made on how best to *represent* our data, both in terms of their spatial properties and the various attribute values which we need to store. Data storage and maintenance is discussed at length in Chapter 3, and we will not go into further detail here. We will from here on assume that the acquired data has been put in digital form, that is, it has been converted into computer-readable format, so that we can begin our analysis.

Data manipulation and analysis

Once the data has been collected and organized in a computer system, we can start analysing it. Here, let us look at what processes were involved in the eventual production of the maps of Figure 1.1. Note that the actual 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 of several years.

The initial (buoy) data 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 an estimated value for a property for some location where we have not measured.

Sample measurements

It appears that the following steps took place for the upper two figures (here we look at SST computations only—WS analysis will have been similarly conducted):

1. For each buoy, the average SST for each month was computed, using the daily SST measurements for that month. This is a simple computation.
2. For each buoy, the monthly average SST was taken together with the geo-

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 area 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 previous to the above steps we had obtained data about average SST for the month of December for a series of years. This too may have been spatially interpolated to obtain a ‘normal situation’ December data set of a fine resolution.

Let us first clarify what is meant by a ‘georeferenced’ list. Data is georeferenced if it is associated with some position on the Earth’s surface, by using a spatial reference system. This can be achieved using (longitude, latitude) coordinates, or by other means that we discuss in Chapter 4. The key issue is that there is some kind of coordinate system as a reference. In our list, we have associated average sea surface temperature observations with spatial locations, and thereby we have georeferenced them.

Georeferenced data

In step 3 above, we mentioned spatial interpolation. To understand this issue, it is important to note that sea surface temperature is a property that occurs everywhere in the ocean, and not only at the buoys where measurements are taken. The buoys only provide a set of sample observations of sea surface temperature. We can use these sample measurements to estimate the value of SST in places where we have not measured it, using a technique called *spatial interpolation*. The theory of spatial interpolation is extensive, but this is not the place to

Spatial interpolation

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.

discuss it. There are in fact many different spatial interpolation techniques, not just one, and some are better in specific situations than others. This is however a typical example of functions that a GIS can perform on user data.

Data presentation

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

Many issues arise in this phase. Among other things, we need to consider what the message is that we want to portray, who the audience is, what kind of presentation medium will be used, which rules of aesthetics apply, and what techniques are available for representation. These issues may sound a little abstract, so let us clarify with the El Niño case.

For Figure 1.1, we can make the following statements:

- The *message* we wanted to portray is what are the El Niño and La Niña events, both in absolute figures, but also 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, (printed matter of A4 size) and possibly 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 north-up; with clear georeferencing; 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 (horizontally) compared to longitude (vertically).

- The *techniques* that we used included the use of a colour scheme and isolines,² plus a number of other techniques.

²Isolines are discussed in Chapter 2.

1.1.3 GISystems, GIScience and GIS applications

The previous discussion defined a *geographic information system*—in the ‘narrow’ sense—in terms of its functions as a computerized system that facilitates the phases of data entry, data management, data analysis and data presentation specifically for dealing with georeferenced data. In the ‘wider’ sense, a functioning GIS requires both hardware and software, and also people such as the database creators or administrators, analysts who work with the software, and the users of the end product. For the purposes of this book we will concern ourselves with the ‘narrow’ definition, and focus on the specifics of these so-called *GISystems*.

GISystems

The discipline that deals with all aspects of the handling of spatial data and geoinformation is called *geographic information science* (often abbreviated to geo-information science or just *GIScience*).

Geo-Information Science is the scientific field that attempts to integrate different disciplines studying the methods and techniques of handling spatial information.

Related terms include *geoinformatics*, *geomatics*, and *spatial information science*. These are all similar terms which have much the same meaning, although each approach has slight differences in the way it deals with problems, some emphasizing engineering approaches, others computational solutions, and so on.

GIScience

As well as being aware of these differences, it is also important to be aware of the difference between a geographic information system and a *GIS application*. In the example discussed above (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 also be used to analyse forest plots in northern Thailand, for instance. That would be a different application, but would make use of the same software. GIS software 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'.

GIS applications

Project-based GIS applications usually have a clear-cut purpose, and these applications can be short-lived: the research is carried out by collecting data, entering data 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.

1.1.4 Spatial data and geoinformation

A 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 confusing their meanings. Occasionally, however, we need to be precise about exactly what it is we are referring to, and in this situation their distinction does matter.

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, such as (x, y) co-ordinates. Sometimes the more precise phrase *geospatial data* is used as a further refinement, which refers to spatial data that is georeferenced. 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. *Geoinformation* is a specific type of information resulting from the interpretation of spatial data.

Geospatial data and
geoinformation

As this information is intended to reduce uncertainty in decision-making, any errors and uncertainties in spatial information products may have practical, financial and even legal implications for the user. For these reasons, it is important that those involved in the acquisition and processing of spatial data are able to assess the *quality* of the base data and the derived information products. 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

Data quality considerations

be used. For this reason, quality is often defined as 'fitness for use'.

Traditionally, most spatial data were collected and held by individual, specialized organizations. In recent years, increasing availability and decreasing cost of data capture equipment has resulted in many users collecting their own data. However, the collection and maintenance of 'base' data remain the responsibility of the various governmental agencies, such as National Mapping Agencies (NMAs), which are responsible for collecting topographic data for the entire country following pre-set standards. Other agencies such as geological survey companies, energy supply companies, local government departments, and many others, all collect and maintain spatial data for their own particular purposes. If 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*.

Base data, sharing and metadata

Since the real power of GIS lies in their ability to combine and analyse georeferenced data from a range of sources, we must pay attention to the issues of data quality and error, as data from different sources are also likely to contain different kinds of error. This may include mistakes or variation in the measurement of position and/or elevation, in the quantitative measurement of attributes or in the labelling or classification of features. Some degree of error is present in every spatial data set. It is important, however, to distinguish between gross errors (blunders or mistakes), which must be detected and removed before the data is used, variations in the data caused by unavoidable measurement and classification errors.

Error in spatial data

It is possible to make a further distinction 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, and a variety of tried and tested techniques is available to describe and evaluate them (see Section 5.2).

Key components of spatial data quality include *positional accuracy* (both horizontal and vertical), *temporal accuracy* (that the data is up to date), *attribute accuracy* (e.g. in labelling of features or of classifications), *lineage* (history of the data including sources), *completeness* (if the data set represents all related features of reality), and *logical consistency* (that the data is logically structured).

Data quality parameters

These components play an important role in assessment of data quality 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.

1.2 The real world and representations of it

One of the main uses of GIS is as a tool to help us make decisions. Specifically, we often want to know the best location for a new facility, the most likely sites for mosquito habitat, or perhaps identify areas with a high risk of flooding so that we can formulate the best policy for prevention. In using GIS to help make these decisions, we need to represent some part of the real world as it is, as it was, or perhaps as we think it will be. We need to restrict ourselves to ‘some part’ of the real world simply because it cannot be represented completely.

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 does not need to 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 can only represent parts of the real world 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 will be based upon the information which we provide the system with. Often, we are dealing with processes or phenomena that change rapidly, or which are difficult to quantify in order to be stored in a computer. It follows that the ways we collect, organise and structure data from the real world plays a key part in this process.

If we have done our job properly, a computer representation of some part of the real world, will allow us to enter and store data, analyse the data and transfer it to humans or to other systems.

1.2.1 Models and modelling

'Modelling' is a term used in many different ways and which has many different meanings. A representation of some part of the real world can be considered a *model* because the representation will have certain characteristics in common with the real world. Specifically, those which we have identified in our model design. This then allows us to study and operate on the model itself instead of the real world in order to test what happens under various conditions, and help us answer 'what if' questions. We can change the data or alter the parameters of the model, and investigate 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 most common, but digital maps also exist, as we shall see in Chapter 7. We will look more closely at maps below. Databases are another important class of models. A database can store a considerable amount of data, and also provides various functions to operate on the stored data. The collection of stored data represents some real world phenomena, so it too is a model. Obviously, here we are especially interested in databases that store spatial data. Digital models (as in a database or GIS) have enormous advantages over paper models (such as maps). They are more flexible, and therefore more easily changed for the purpose at hand. In principle, they allow animations and simulations to be carried out by the computer system. This has opened up an important toolbox that can help to improve our understanding of the world.

Models as representations

The attentive reader will have noted our threefold use of the word 'model'. This, perhaps, may be confusing. Except as a verb, where it 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 that we can observe in reality, usually to enable some type of study, administration, computation and/or simulation. In this book we will use the term *application models* to refer to models with a specific application, including real-world models and so-called analytical models. 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 kinds of data. We discuss these issues further in Chapter 3.

Application models

Most 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 sometime in the future. Dynamic models are inherently more complicated than static models, and usually require much more computation. Simulation models are an important class of dynamic models that allow the simulation of real world processes.

Dynamic models

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 form of dynamic model: a series of 'static snapshots' allowing us to infer some information about the behaviour of the system over time. We will return to modelling issues in Chapter 6.

1.2.2 Maps

As noted above, maps are perhaps the best known (conventional) models of the real world. Maps have been used for thousands of years to represent information about the real world, and continue to be extremely useful for many applications in various domains. Their conception and design has developed into a science with a high degree of sophistication. A disadvantage of the traditional paper map is that it is generally restricted to two-dimensional static representations, and that it is always displayed in a fixed 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. Hence, the selection of a proper map scale is one of the first and most important steps in map design.

Map scale and accuracy

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 applications, including the development of other maps.

Cartography

With the advent of computer systems, analogue cartography developed into digital cartography, and computers play an integral part in modern cartography. Alongside this trend, the role of the map has also changed accordingly, and the dominance of paper maps is eroding in today's increasingly 'digital' world. The traditional role of paper maps as a data storage medium is being taken over

Digital maps

by (spatial) databases, which offer a number of advantages over 'static' maps, as discussed in the sections that follow. Notwithstanding these developments, paper maps remain as important tools for the display of spatial information for many applications.

1.2.3 Databases

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

1. A database can be used by multiple users at the same time—i.e. it allows *concurrent use*,
2. A database offers a number of techniques for storing data and allows the use of the most efficient one—i.e. it supports *storage optimization*,
3. A database allows the imposition of rules on the stored data; rules that will be automatically checked after each update to the data—i.e. it supports *data integrity*,
4. A database offers an easy to use data manipulation language, which allows the execution of all sorts of data extraction and data updates—i.e. it has a *query facility*,
5. A 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 kind of data. Modern database systems, as we shall see in Section 3.4, 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 for a table to have 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 entire El Niño buoy measurements database is likely to have more tables than the one illustrated. There may be data available about the buoys' maintenance and service schedules; there may also 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 a query language. A query was defined that computes the monthly average SST from the daily measurements, for each buoy. A discussion of the particular query language that was used is outside the scope of this book, but we should mention that the query was a simple program with just four lines of code.

1.2.4 Spatial databases and spatial analysis

A GIS must store its data in some way. For this purpose the previous generation of software was equipped with relatively rudimentary facilities. Since the 1990's there has been an increasing trend in GIS applications that used a GIS for spatial analysis, and used a database for storage. In more recent years, *spatial databases* (also known as geodatabases) have emerged. Besides traditional administrative data, they can store representations of real world geographic phenomena for use in a GIS. These databases are special because they use additional techniques different from tables to store these spatial representations.

A geodatabase is not the same thing as a GIS, though both systems share a number of characteristics. These include the functions listed above for databases in general: concurrency, storage, integrity, and querying, specifically, but not only, spatial data. A GIS, on the other hand, is tailored to operate on *spatial* data. It 'knows' about spatial reference systems, and supports all kinds of analyses that are inherently geographic in nature, such as distance and area computations and spatial interpolation. This is probably GIS's main strength: providing various ways to combine representations of geographic phenomena. GISs, moreover, built-in tools for map production, of the paper and the digital kind. They operate with an 'embedded understanding' of geographic space. Databases typically lack this kind of understanding.

Geodatabases

The phenomena for which we want to store representations in a spatial database may have point, line, area or image characteristics. Different storage techniques exist for each of these kinds of spatial data.³ These geographic phenomena have various relationships with each other and possess spatial (geometric), thematic

Representations of
geographic phenomena

³Since we also have different analytical techniques for these different types of data, an im-

and temporal attributes (they exist in space and time). For data management purposes, phenomena are classified into thematic data layers. The purpose of the database is usually described by a description such as cadastral, topographic, land use, or soil database.

Spatial analysis is the generic term for all manipulations of spatial data carried out to improve one's understanding of the geographic phenomena that the data represents. It involves questions about how the data in various layers might relate to each other, and how it varies over space. For example, in the El Niño case, we may want to identify the steepest gradient in water temperature. The aim of spatial analysis is usually to gain a better understanding of geographic phenomena through discovering patterns that were previously unknown to us, or to build arguments on which to base important decisions. It should be noted that some GIS functions for spatial analysis are simple and easy-to-use, others are much more sophisticated, and demand higher levels of analytical and operating skills. Successful spatial analysis requires appropriate software, hardware, and perhaps most importantly, a competent user.

Spatial analysis

portant choice in the design of a spatial database application is whether some geographic phenomenon is better represented as a point, as a line, or as an area. Currently, spatial databases support the storage of image data, but that support still remains relatively limited.

1.3 Structure of this book

This chapter has attempted to provide a 'gentle' introduction to GIS. It has discussed the nature of GIS tools and GIS as a field of scientific research. Much of the technical detail has been intentionally left out in favour of a broader discussion of the key issues relating to both of these topics. The chapter has looked at the purposes of GIS and identified understanding objects and events in geographic space as the common thread amongst GIS applications, and that spatial data and spatial data processing are key factors in this understanding. A simple example of a study of the EL Niño effect provided an illustration, without the technical details.

It was noted that the use of GIS commonly takes place in several phases: data capture and preparation, storage and maintenance, manipulation and analysis, and data presentation. Before we get to discussing these phases, the following two chapters provide more discussion on important background concepts and issues. In Chapter 2, we will focus the discussion on different kinds of geographic phenomena and their representation in a GIS, and discuss appropriate instances of when to use which. Chapter 3 is devoted to a discussion of data processing systems for spatial data, namely, GIS, databases and spatial databases.

Following these last two chapters, the remaining structure of the book follows the phases identified above. In Chapter 5 we look 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. Analysis of geoinformation is the focus of Chapter 6. It discusses the most important forms of spatial data analysis in some detail, and looks at issues related to spatial modelling.

The phase of data *visualization* is the topic of Chapter 7. This chapter deals with fundamental cartographic principles: what to put on a map, where to put it, and what techniques to use for specific types of data. Sooner or later, almost all GIS users will be involved the presentation of geoinformation (usually of maps), so it is important to understand the underlying principles.

Questions

1. Take another look at the list of professions provided on page 26. 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 *a combination of the two*. 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 37, 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. But often, useful data has not been converted in this way. From your own experience, provide examples of data sources in non-digital format.
6. Assume the El Niño project is operating with just four buoys, and not 70, and their location is as illustrated in Figure 1.4. We have already computed

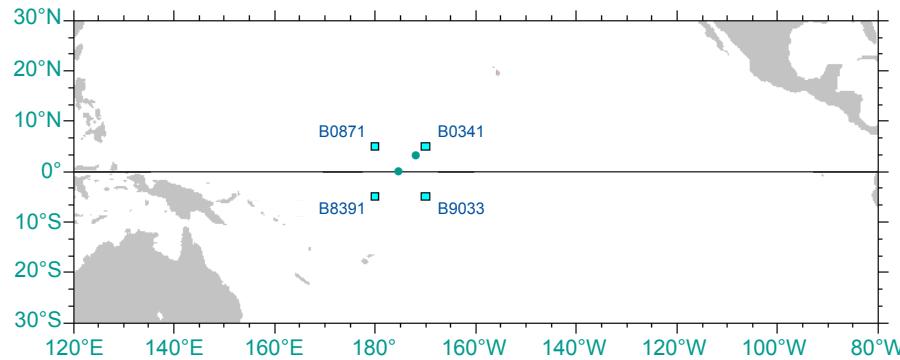


Figure 1.4: Just four measuring buoys

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. In Table 1.2, we illustrated some stored measurement data. The table uses one row of data for a single day that some 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 these locations are stored?

Chapter 2

Geographic information and Spatial data types

[previous](#)

[next](#)

[back](#)

[exit](#)

[contents](#)

[index](#)

[glossary](#)

[web links](#)

[bibliography](#)

[about](#)

2.1 Models and representations of the real world

As discussed in the previous chapter, we use GISs to help analyse and understand more about processes and phenomena in the *real world*. Section 1.2.1 referred to the process of *modelling*, or building a representation which has certain characteristics in common with the real world. In practical terms, this refers to the process of representing key aspects of the real world digitally (inside a computer). These representations are made up of spatial data, stored in memory in the form of bits and bytes, on media such as the hard drive of a computer. This digital representation can then be subjected to various analytical functions (computations) in the GIS, and the output can be visualized in various ways.

Modelling is the process of producing an abstraction of the ‘real world’ so that some part of it can be more easily handled.

Depending on the application domain of the model, it may be necessary to manipulate the data with specific techniques. To investigate the geology of an area, we may be interested in obtaining a geological classification. This may result in additional computer representations, again stored in bits and bytes. To examine how the data is stored inside the GIS, one could look into the actual data files, but this information is largely meaningless to a normal user.

As highlighted in Figure 2.1, the process of translating the relevant aspects of the real world into a computer representation of it is a domain of expertise by itself. It might be achieved through direct observations using sensors, and digitizing (converting) the sensor output for computer usage. This is the domain of *remote sensing*, the topic of *Principles of Remote Sensing* [53]. We may also do

this by indirect means: for instance, by making use of the output of a previous project, such as a paper map, and re-digitizing it.

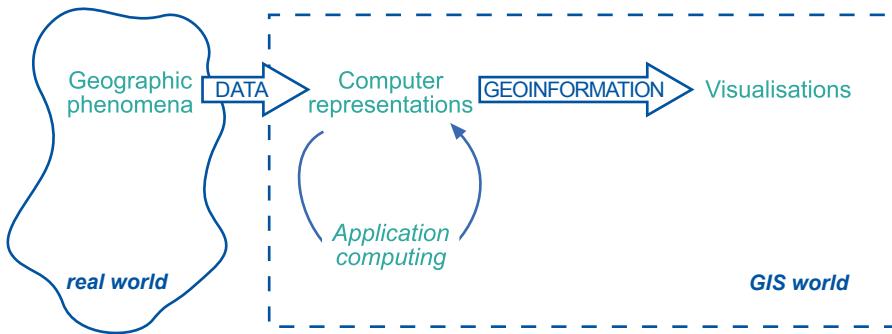


Figure 2.1: Representing relevant aspects of real-world phenomena inside a GIS to build models or simulations.

In order to better understand both our representation of the phenomena, and our eventual output from any analysis, we can use the GIS to create *visualizations* from the computer representation, either on-screen, printed on paper, or otherwise.¹ It is crucial to understand the fundamental differences between these notions. The real world, after all, is a completely different domain than the 'GIS' world, in which we build models or simulations of the real world.

Given the complexity of real world phenomena, our models can by definition never be perfect. We have limitations on the amount of data that we can store, limits on the amount of detail we can capture, and (usually) limits on the time we have available for a project. It is therefore possible that some facts or relation-

Complexity

¹It should be mentioned here 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. The map-like illustrations in this chapter purposely do not have a legend or text tags. They are *not* intended to be maps.

ships that exist in the real world may not be discovered through our ‘models’.

Any geographic phenomenon can usually be represented in various ways; the choice of which representation is best depends mostly on two issues. Firstly, what original, raw data (from sensors or otherwise) is available, and secondly, what sort of data manipulation is required or will be undertaken. Key aspects of data acquisition and preparation are discussed in Chapter 5. This chapter will examine various types of geographic phenomena in more depth, and the types of computer representations available for them.

Representation of
phenomena

2.2 Geographic phenomena

2.2.1 Defining geographic phenomena

A GIS operates under the assumption that the relevant spatial phenomena occur in a two- or three-dimensional *Euclidean space*, unless otherwise specified. Euclidean space can be informally defined as a model of space in which locations are represented by coordinates— (x, y) in 2D; (x, y, z) in 3D—and *distance* and *direction* can be defined with geometric formulas. In the 2D case, this is known as the *Euclidean plane*, which is the most common Euclidean space in GIS use.

Euclidean space

In order to be able to represent relevant aspects of real world phenomena inside a GIS, we first need to define what it is we are referring to. We might define a geographic phenomenon as a manifestation of an entity or process of interest that:

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

The relevant phenomena for a given application depends entirely on one's objectives. For instance, in water management, the objects of study might 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. Note that all of these can be named or described, georeferenced and provided with a time interval at which each exists. In multipurpose cadastral administration, the objects of study are different: houses, land parcels, streets of various types, land use forms, sewage canals and other

Objectives of the application

forms of urban infrastructure may all play a role. Again, these can be named or described, georeferenced and assigned a time interval of existence.

Not all relevant phenomena come as triplets (*description, georeference, 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 not considered relevant. If the time interval is missing, we might 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, then we have something that exists in space and time, yet cannot be described. Obviously this last issue very much limits the usefulness of the information.

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.

2.2.2 Types of geographic phenomena

The attempted definition of geographic phenomena above is necessarily abstract, and therefore perhaps somewhat difficult to grasp. The main reason for this is that geographic phenomena come in so many different ‘flavours’, which we will try to categorize below. Before doing so, we must make two further observations.

Firstly, In order to be able to represent a phenomenon in a GIS requires us to state *what* it is, and *where* it is. We must provide a description—or at least a name—on the one hand, and a georeference on the other hand. We will skip over the temporal issues for now, and come back to these in Section 2.5. The reason for this is that current GISs do not provide much automatic support for time-dependent data, and that this topic must be therefore be considered an issue of advanced GIS use.

Secondly, some phenomena manifest themselves essentially everywhere in the study area, while others only do so in certain localities. If we define our study area as the equatorial Pacific Ocean, 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*.

Fields

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

Some common examples of geographic fields are air temperature, barometric pressure and elevation. These fields are in fact continuous in nature. Examples of discrete fields are land use and soil classifications. For these too, any location

in the study area is attributed a single land use class or soil class. We discuss fields further in Section 2.2.3.

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*.

Objects

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

A simple rule-of-thumb is that natural geographic phenomena are usually fields, and man-made phenomena are usually 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 2.2.4.

Elevation in the Falset study area, Tarragona province, Spain. The area is approximately 25×20 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.

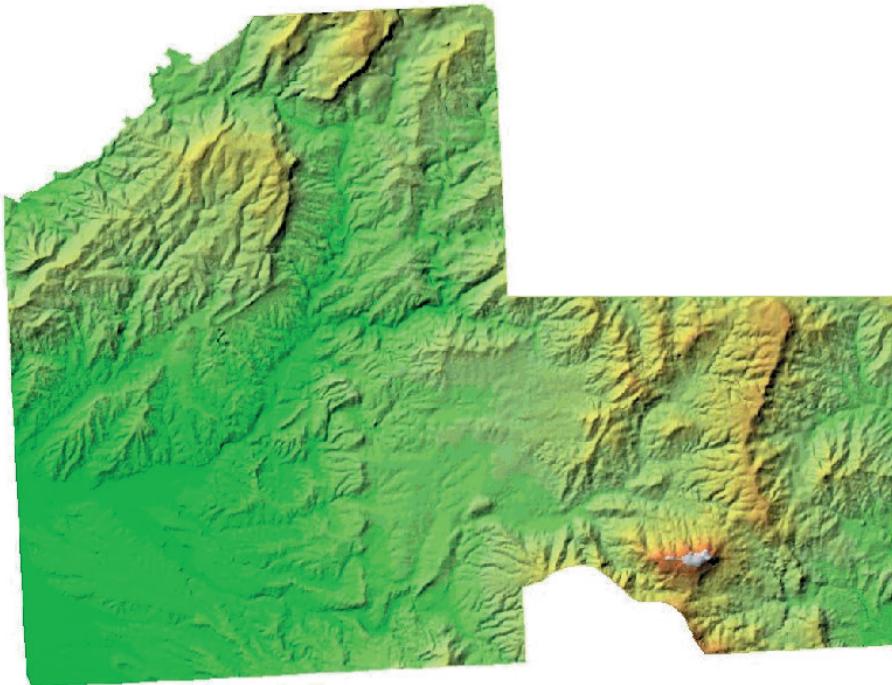


Figure 2.2: A continuous field example, namely the *elevation* in the study area of Falset, Spain.

Data source: Department of Earth Systems Analysis (ESA, ITC)

2.2.3 Geographic fields

A field is a geographic phenomenon that has a value ‘everywhere’ in the study area. We can therefore think of a field as a mathematical function f that associates a specific value with any position in the study area. Hence if (x, y) is a position in the study area, then $f(x, y)$ stands for the value of the field f at locality (x, y) .

Fields can be *discrete* or *continuous*. In a *continuous field*, the underlying function is assumed to be ‘mathematically smooth’, meaning that the field values along any path through the study area do not change abruptly, but only gradually. Good examples of continuous fields are air temperature, barometric pressure, soil salinity and elevation. Continuity means that all changes in field values are gradual. A continuous field can even be *differentiable*, meaning we can determine a measure of change in the field value per unit of distance anywhere and in any direction. For example, if the field is elevation, this measure would be slope, i.e. the change of elevation per metre distance; if the field is soil salinity, it would be salinity gradient, i.e. the change of salinity per metre distance. 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.

Discrete fields divide the study space in mutually exclusive, bounded parts, with all locations in one part having the same field value. Typical examples are land classifications, for instance, using either geological classes, soil type, land use type, crop type or natural vegetation type. An example of a discrete field—in this case 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 as-

Continuous fields

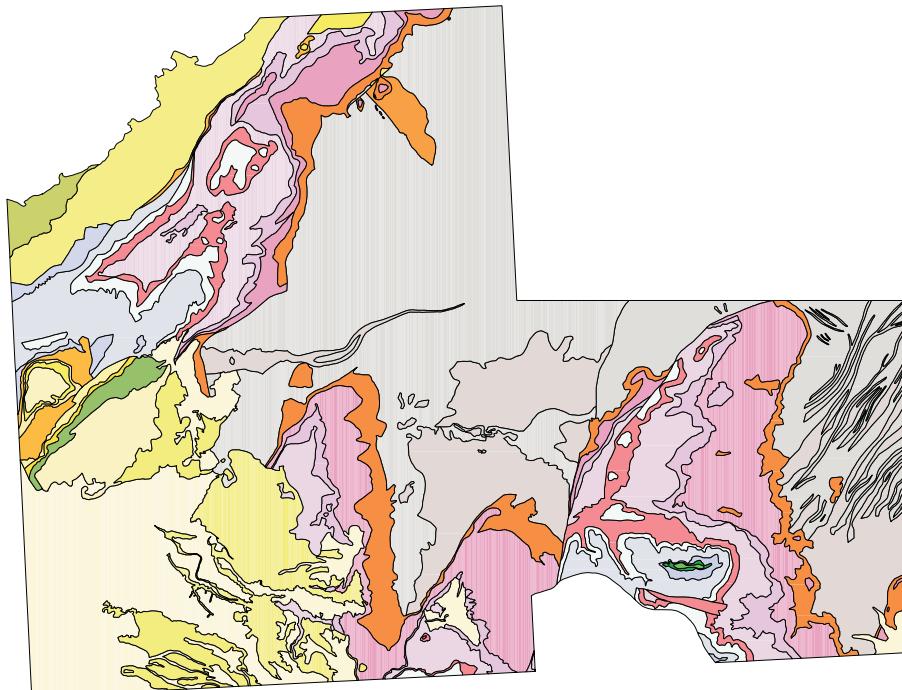
Discrete fields

signed 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.

Essentially, these two types of fields differ in the type of cell values. A discrete field like landuse type will store cell values of the type ‘integer’. Therefore it is also called an integer raster. Discrete fields can be easily converted to polygons, since it is relatively easy to draw a boundary line around a group of cells with the same value. A continuous raster is also called a ‘floating point’ raster. A *field-based model* consists of a finite collection of geographic fields: we may be interested in elevation, barometric pressure, mean annual rainfall, and maximum daily evapotranspiration, and thus use four different fields to model the relevant phenomena within our study area.

Field-based model

- Miocene and Quaternary (lower left)
- Oligocene (left)
- Cretaceous (right)
- Eocene
- Lias
- Keuper and Muschelkalk
- Buntsandstein
- Intrusive and sedimentary areas



Observe that—typical for fields—with any location only a single geological unit is associated. As this is a *discrete* field, value changes are discontinuous, and therefore locations on the boundary between two units are *not* associated with a particular value (i.e. with a geological unit).

Figure 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: Department of Earth Systems Analysis (ESA, ITC)

Data types and values

Since we have now differentiated between continuous and discrete fields, we may also look at different kinds of data values which we can use to represent our ‘phenomena’. It is important to note that some of these data types limit the types of analyses that we can do on the data itself:

1. *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 called *categorical data* when the values assigned are sorted according to some set of non-overlapping categories. For example, we might identify the soil type of a given area to belong to a certain (pre-defined) category.
2. *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.
3. *Interval data values* are quantitative, in that they allow simple forms of computation like addition and subtraction. However, interval data has 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.
4. *Ratio data values* allow most, if not all, forms of arithmetic computation.

Rational data have a natural zero value, and multiplication and division of values are possible operators (distances measured in metres are an example). Continuous fields can be expected to have ratio data values, and hence we can interpolate them.

We usually refer to nominal and categorical data values as 'qualitative' data, because we are limited in terms of the computations we can do on this type of data. Interval and ratio data is known as 'quantitative' data, as it refers to quantities. However, ordinal data does not seem to fit either of these data types. Often, ordinal data refers to a ranking scheme or some kind of hierarchical phenomena. Road networks, for example, are made up of motorways, main roads, and residential streets. We might expect roads classified as motorways to have more lanes and carry more traffic than a residential street.

Qualitative and quantitative data

2.2.4 Geographic objects

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

- *Location* (where is it?),
- *Shape* (what form is it?),
- *Size* (how big is it?), and
- *Orientation* (in which direction is it facing?).

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 an in-car navigation system, all that matters about geographic objects like petrol stations is where they are. Thus, location alone is enough to describe them in this particular context, and shape, size and orientation are not necessarily relevant. 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*. This relates to whether an object is perceived as a point feature, or a linear, area or volume feature. 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.

Dimensionality of features

Figure 2.4 illustrates geological faults in the Falset study area, a typical example of a geographic phenomenon that is made up of objects. Each of the faults has a location, and here the fault's shape is represented 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 more often we look at *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 an in-car navigation system, the collection of roads in that system, the collection of land parcels in a cadastral system, and in many more cases. We will see in Section 2.3 that this natural law of ‘mutual non-overlap’ has been a guiding principle in the design of computer representations of geographic phenomena.

Collections of geographic objects can be interesting phenomena at a higher aggregation level: forest plots form forests, groups of parcels form suburbs, streams, brooks and rivers form a river drainage system, roads form a road network, and SST buoys form an SST sensor network. It is sometimes useful to view geographic phenomena at this more aggregated level and look at characteristics like

Geographic scale

coverage, connectedness, and capacity. For example:

- 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)



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: Department of Earth Systems Analysis (ITC)

- How many cars can optimally travel from one city to another in an hour?
(A capacity question)

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, discussed in Section 2.3.4.

2.2.5 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.

Crisp and fuzzy boundaries

As a general rule-of-thumb, 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 there is still limited support for these in existing GIS software. The areas identified in a geological classification, like that of Figure 2.3, are typically vaguely bounded in reality, but applications of this geological information probably do not require high positional accuracy of the boundaries involved. Therefore, an assumption that they are actually crisp boundaries will have little influence on the usefulness of the data.

2.3 Computer representations of geographic information

Up to this point, we have not looked at how geoinformation, like fields and objects, is represented in a computer. After the discussion of the main characteristics of geographic phenomena above, let us now examine representation in more detail. We have seen that various geographic phenomena have the characteristics of continuous functions over space. Elevation, for instance, can be measured at many locations, even within one's own backyard, and each location may give a different value. In order to represent such a phenomenon faithfully in computer memory, we could either:

- Try to store as many (*location, elevation*) observation pairs as possible, or
- Try to find a symbolic representation of the elevation field function, as a formula in x and y —like $(3.0678x^2 + 20.08x - 7.34y)$ or so—which can be evaluated to give us the elevation at any given (x, y) location.

Both of these 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 do not know just what this function should look like, and that it would be extremely difficult to derive such a function for larger areas. In GISs, typically a combination of both approaches is taken. We store a finite, but intelligently chosen set of (sample) locations with their elevation. This gives us the elevation for those stored locations, but not for others. We can use an interpolation function that allows us to infer a reasonable elevation value for locations

Interpolating sample values

that are not stored. A simple and commonly used interpolation function 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. Interpolation of point data discussed in more detail in Section 5.4.

Interpolation is made possible by a principle called *spatial autocorrelation*. This is a fundamental principle which refers to the fact that locations that are closer together are more likely to have similar values than locations that are far apart—commonly referred to as ‘Tobler’s first law of Geography’. An obvious example of a phenomenon which exhibits this property is sea-surface temperature, where one might expect a high degree of correlation between measures taken close together (refer to the SST example of Chapter 1).

Spatial autocorrelation

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 are often erratically curved. A famous paradoxical question is whether one can actually measure the length of Great Britain’s coastline, i.e. can one measure around rocks, pebbles or even grains of sand?² In a computer, such random, curvilinear features can never be fully represented, and usually require some degree of generalization.

Boundaries

From this it becomes clear that phenomena with intrinsic continuous and/or infinite characteristics have to be represented with finite means (computer memory) for computer manipulation, and any finite representation scheme is open to errors of interpretation. To this end, fields are usually implemented with a

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

tessellation approach, and objects with a (topological) *vector* approach. however, this is not a hard-and-fast rule, as practice sometimes demands otherwise.

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

2.3.1 Regular tessellations

A *tessellation* (or tiling) is a partitioning 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).

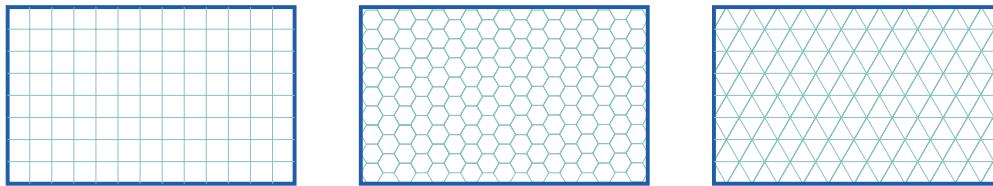


Figure 2.5: The three most common regular tessellation types: square cells, hexagonal cells, and triangular cells.

In all regular tessellations, the cells are of the same shape and size, and 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. These tessellations are known under various names in different GIS packages, but most frequently as *rasters*.

A *raster* is a set of regularly spaced (and contiguous) cells with associated (field) values. The associated values represent *cell* values, not point values. This means that the value for a cell is assumed to be valid for all locations within the cell.

The size of the area that a single raster cell represents is called the raster's *resolution*. Sometimes, the word *grid* is also used, but strictly speaking, a *grid* refers to values at the intersections of a network of regularly spaced horizontal and perpendicular lines (see Figure 2.6). Grids are often used for discrete measurements that occur at regular intervals. Grid values are often considered synonymous with raster cells, although they are not.

Grids and rasters

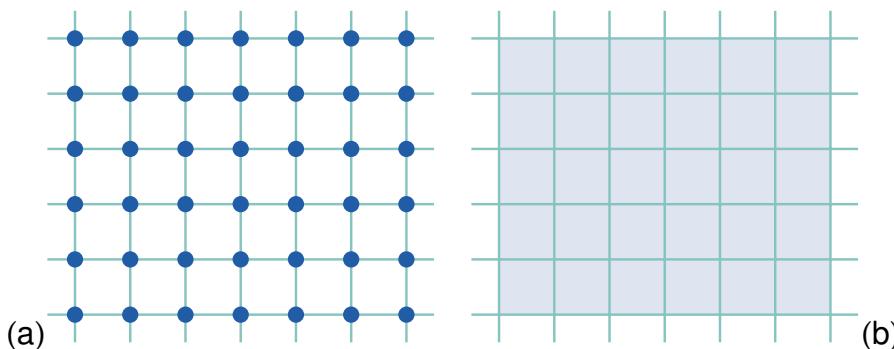


Figure 2.6: A grid (a) is a collection of regularly spaced (field) values, while a raster (b) is composed of cells. The associated values with each grid point or raster cell are not illustrated.

There are some issues related to cell-based partitioning of the study space. The field value of a cell can be interpreted as one for the complete tessellation cell, in which case the field is discrete, not continuous or even differentiable. Some convention is needed to state which value prevails on cell boundaries; with square cells, this convention 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 intensive 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 us to represent continuous, even differentiable, functions.

An important advantage of regular tessellations is that we 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 phenomena of interest. For example, suppose we use any of the above regular tessellations to represent elevation in a perfectly flat area. In this case we need just as many cells as in a strongly undulating terrain: the data structure does not adapt to the lack of relief. We would, for instance, still use the $m \times n$ cells for the raster, although the elevation might be 1500 m above sea level everywhere.

2.3.2 Irregular tessellations

Above, we discussed that regular tessellations provide simple structures with straightforward algorithms, which are, however, not adaptive to the phenomena they represent. Essentially this means they might not represent the phenomena in the most efficient way. For this reason, substantial research 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 also been implemented.

Irregular tessellations are adaptive

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 2.7. 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

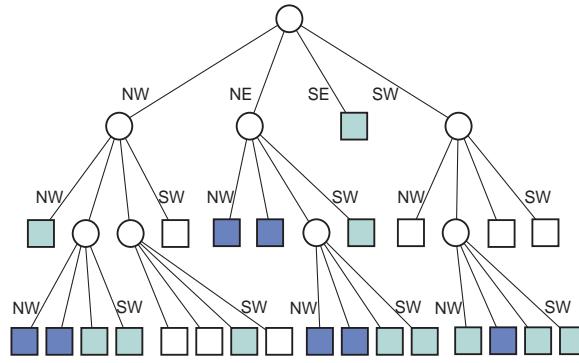
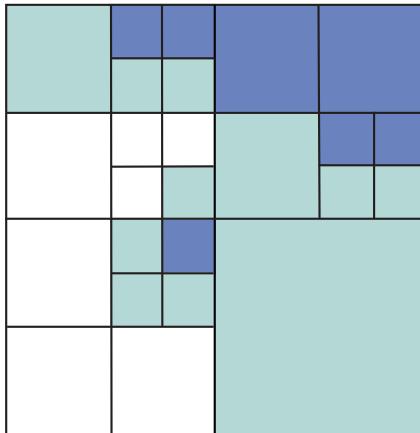


Figure 2.7: An 8×8 , three-valued raster (here: colours) and its representation as a region quadtree. To construct the quadtree, the field is successively split into four quadrants until parts have only a single field value. After the first split, the southeast quadrant is entirely green, and this is indicated by a green square at level two of the tree. Other quadrants had to be split further.

Quadtrees are adaptive because they apply the *spatial autocorrelation* principle, i.e. that 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. The square nodes at the same level represent equal area sizes, allowing quick computation of the area associated with some field value. The top node of the tree represents the complete raster.

To summarise the above discussion, we can say that tessellations partition the

study space into cells, and assign a value to each cell. A raster is a regular tessellation with square cells (by far the most commonly used). The method by which the study space is split into cells is (to some degree) arbitrary, as cell boundaries usually have little or no bearing to the real world phenomena that are represented.

2.3.3 Vector representations

Tessellations do not explicitly store georeferences of the phenomena they represent. Instead, they provide a georeference of the lower left corner of the raster, for instance, plus an indicator of the raster's resolution, thereby implicitly providing georeferences for all cells in the raster. In *vector representations*, an attempt is made to explicitly associate georeferences with the geographic phenomena. A georeference is a coordinate pair from some geographic space, and is also known as a vector. This explains the name. 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.

Vectors store georeferences
explicitly

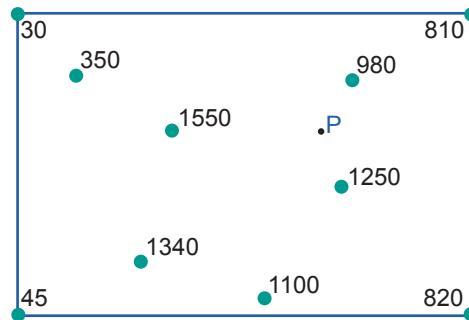


Figure 2.8: Input locations and their (elevation) values for a TIN construction. The location P is an arbitrary location that has no associated elevation measurement.

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 2.8. From these 3D points, we can construct an irregular tessellation made of triangles. Two such tessellations are illustrated in Figure 2.9.

TINs represent a continuous field

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

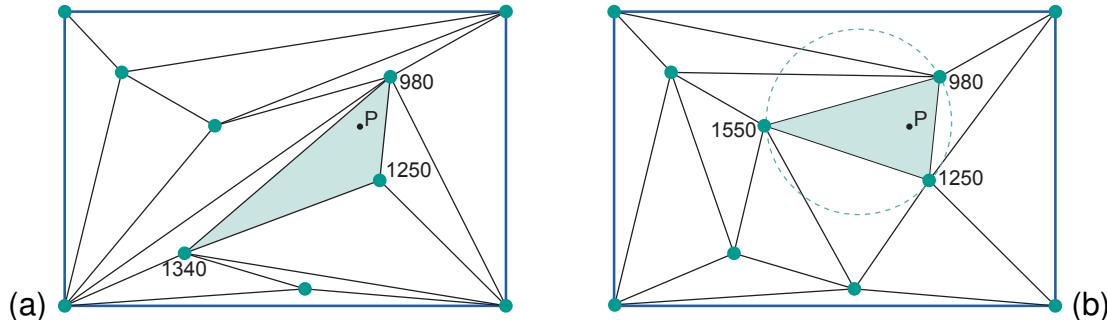


Figure 2.9: Two triangulations based on the input locations of Figure 2.8. (a) one with many ‘stretched’ triangles; (b) the triangles are more equilateral; this is a *Delaunay triangulation*.

to compute an approximation of elevation of other locations.³ Since we can pick many triples of points, we can construct many such planes, and therefore we can have many elevation approximations for a single location, such as P (Figure 2.8). 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 2.9 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

³*Slope* is usually defined to consist of two parts: the *gradient* and the *aspect*. 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 6.4.4.

P to the three triangle anchors is smaller.

The triangulation of Figure 2.9(b) happens to be a *Delaunay triangulation*, which in a sense is an optimal triangulation. There are multiple ways of defining what such a triangulation is (see [46]), 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 of Figure 2.9(b).

Delaunay triangulation

A TIN clearly is a vector representation: each anchor point has a stored georeference. Yet, we might also call it an irregular tessellation, as the chosen triangulation provides a partitioning of the entire study space. However, in this case, the cells do not have an associated stored value as is typical of tessellations, but rather a simple interpolation function that uses the elevation values of its three anchor points.

Point representations

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

Points are used to represent objects that are best described as shape- and size-less, one-dimensional 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, a park will not usually be considered a point feature, but perhaps a museum will, and certainly a public phone booth might be represented as a point.

Besides the georeference, usually extra data is stored for each point object. This so-called *attribute* 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, or the data last serviced.

Line representations

Line data are used to represent one-dimensional objects such as roads, railroads, canals, rivers and power lines. Again, there is an issue of relevance for the application and the scale that the application requires. For the example 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.

Above, we discussed the notion that arbitrary, continuous curvilinear features are as equally difficult to represent as continuous fields. GISs therefore approximate such features (finitely!) as lists of *nodes*. The two *end nodes* and zero or more *internal nodes* or *vertices* define a *line*. Other terms for 'line' that are commonly used in some GISs are polyline, arc or edge. A node or vertex is like a point (as discussed above) but it only serves to define the line, and provide shape in order to obtain a better approximation of the actual feature.

Nodes and vertices

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, but many systems do not at present accommodate such shapes. If a GIS supports some of these curvilinear features, it does so using parameterized mathematical descrip-

Representing curved lines

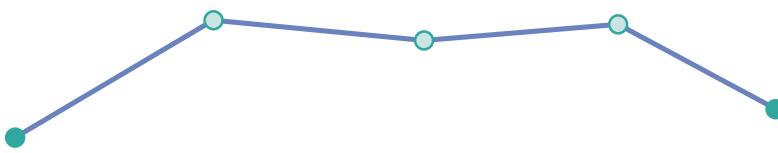


Figure 2.10: A line is defined by its two end nodes and zero or more internal nodes, also known as vertices. This line representation has three vertices, and therefore four line segments.

tions. 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 types of interesting questions arise that have to do with connectivity and network capacity. These relate to applications such as traffic monitoring and watershed management. With network elements—i.e. the lines that make up the network—extra values are commonly associated like distance, quality of the link, or carrying capacity.

Networks

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.

Polygons

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 2.11. 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.

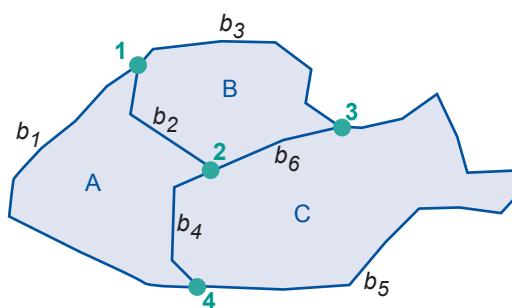


Figure 2.11: Areas as they are represented by their boundaries. Each boundary is a cyclic sequence of line features; each line—as before—is a sequence of two end nodes, with in between, zero or more vertices.

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 2.11. The line that makes up the boundary between them is the same, which means that using the above representation the line would be stored twice, namely once for each polygon. This is a form of data duplication—known as *data redundancy*—which is (at least in theory,) unnecessary, although it remains a feature of some systems.

Data redundancy

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 2.11, 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 neighbouring 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 2.12. It illustrates which additional information is stored about spatial relationships between lines and polygons. Obviously, real coordinates for nodes (and vertices) will also be stored in another table.

Boundary model

The boundary model is sometimes also called the *topological data model* as it cap-

<i>line</i>	<i>from</i>	<i>to</i>	<i>left</i>	<i>right</i>	<i>vertexlist</i>
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	...

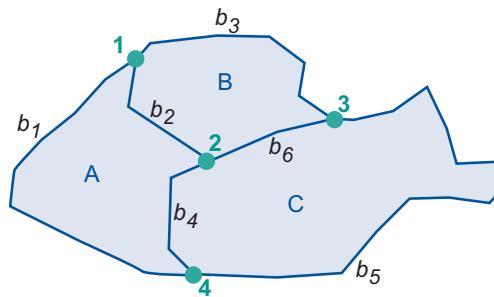


Figure 2.12: A simple boundary model for the polygons A, B and C. For each arc, we store the start and end node (as well as a vertex list, but these have been omitted from the table), its left and right polygon. The ‘polygon’ W denotes the outside world polygon.

tures some topological information, such as polygon neighbourhood. Observe that it is a simple query to find all the polygons that are the neighbour of some given polygon, unlike the case above.

2.3.4 Topology and spatial relationships

General spatial topology

Topology deals with spatial properties that do not change under certain transformations. For example, features drawn on a sheet of rubber (as in Figure 2.13) can be made to change in shape and size by stretching and pulling the sheet. However, some properties of these features 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 perimeters have changed.

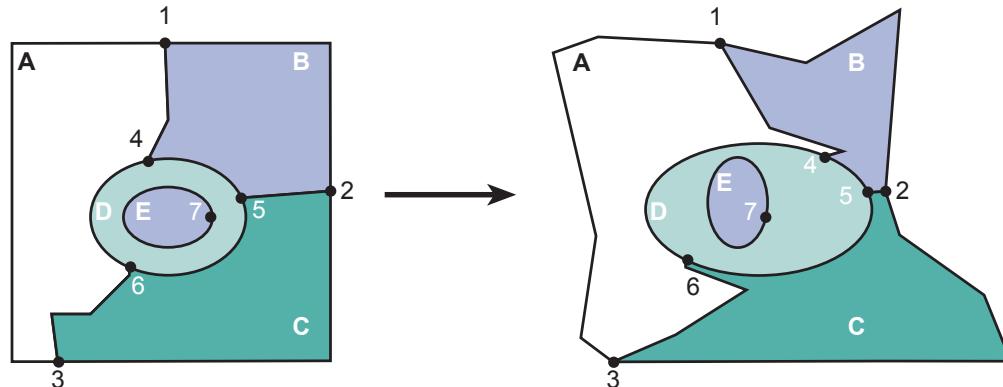


Figure 2.13: Rubber sheet transformation: The space is transformed, yet many relationships between the constituents remain unchanged.

Topology refers to the spatial relationships between geographical elements in a data set that do not change under a continuous transformation.

Topological relationships are built from simple elements into more complex elements: nodes define line segments, and line segments connect to define lines, which in turn define polygons. The fundamental issues relating to order, connectivity and adjacency of geographical elements form the basis of more sophisticated GIS analyses. These relationships (called topological properties) are invariant under a continuous transformation, referred to as a topological mapping.

Topological properties

In what follows below, we will look at aspects of topology in two ways. Firstly, using *simplices*, we will look at how simple elements (points) can be combined to define more complex ones (lines and polygons). Secondly, we will examine the logical aspects of topological relationships using *set-theory*. The three-dimensional case is also briefly discussed.

Topological relationships

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

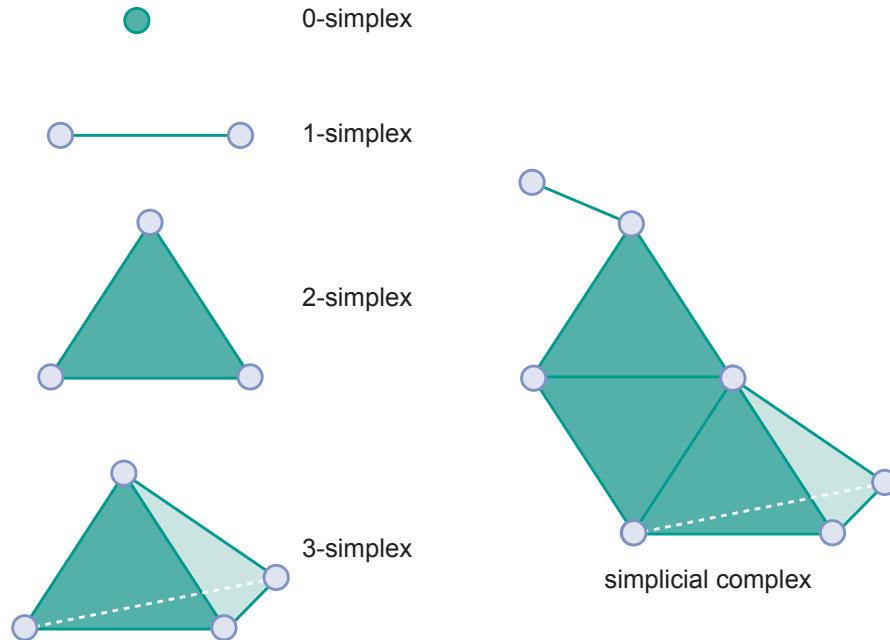


Figure 2.14: Simplices and a simplicial complex. Features are approximated by a set of points, line segments, triangles, and tetrahedrons.

'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 2.14 provides examples.

As the topological characteristics of simplices are well-known, we can infer the topological characteristics of a simplicial complex from the way it was constructed.

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 largest set of points of 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.

Interior and exterior

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 $\text{boundary}(A)$ and $\text{interior}(A)$, 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 using set theory, as

Set theory

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

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

In the above formula, the symbol \wedge expresses the logical connective ‘and’. Thus, the formula states four properties that must all hold for the formula to be true.

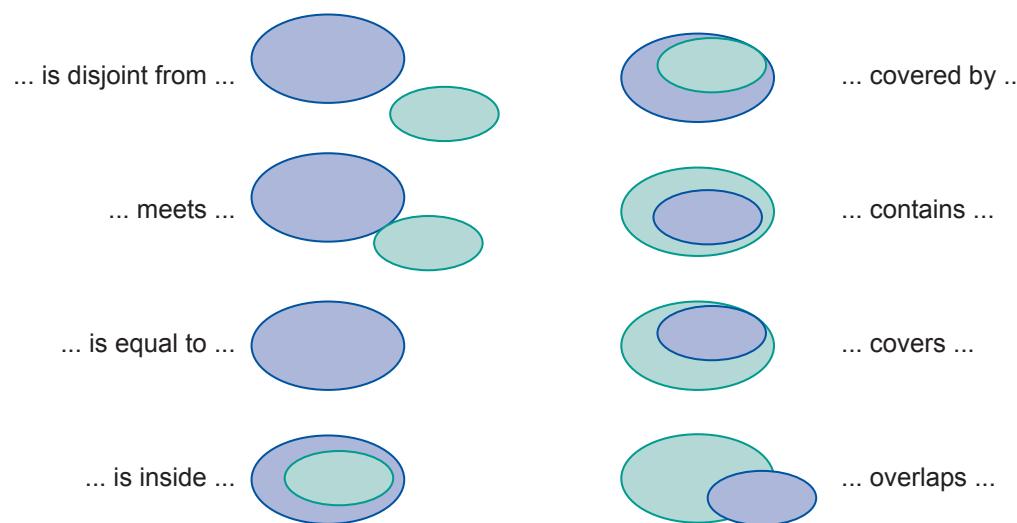


Figure 2.15: Spatial relationships between two regions derived from the topological invariants of intersections of boundary and interior. The relationships can be read with the green region on the left ... and the blue region on the right ...

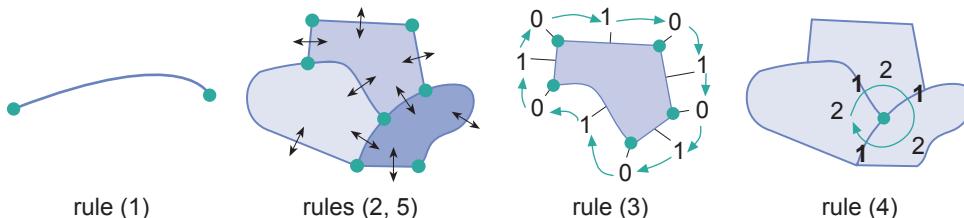
Figure 2.15 shows all eight spatial relationships: *disjoint*, *meets*, *equals*, *inside*, *covered by*, *contains*, *covers*, and *overlaps*. These relationships can be used in queries against a spatial database, and represent the ‘building blocks’ of more complex spatial queries.

It turns out that the rules of how simplices and simplicial complexes can be embedded 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 2.16.

Topological consistency

Figure 2.16: The five rules of topological consistency in two-dimensional space

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.



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 almost purely 2D, and this remains the case for the majority of present-day GIS applications. Many application domains make use of elevational, 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 2.16. 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 -coordinates, but different z -values. Such nodes would constitute a perfectly vertical feature, and this is not allowed. 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 these are worth mentioning here: mineral exploration, where solids are used to represent ore bodies, and urban models, 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 already has six.

2.3.5 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 sometimes defy an easy answer.

Map scale can be defined as the ratio between the distance on a paper map and the distance of the same stretch in the terrain. A 1:50,000 scale map means that 1 cm on the map represents 50,000 cm, i.e. 500 m, in the terrain. ‘Large-scale’ means that the ratio is large, so typically it means there is much detail, as in a 1:1,000 paper map. ‘Small-scale’ in contrast means a small ratio, hence less detail, as in a 1:2,500,000 paper map. When applied to spatial data, the term *resolution* is commonly associated with the cell width of the tessellation applied.

Large-scale and small-scale maps

Digital spatial data, as stored in a GIS, is essentially without scale: scale is a ratio notion associated with visual output, like a map or on-screen display, 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 5.1.2), and in this context we might informally say that the data is at this-or-that scale, indicating the scale of the map from which the data was derived.

When digital spatial data sets have been collected with a specific map-making purpose in mind, and these maps were designed to be of a single map scale, like 1:25,000, we might suppose that the data carries the characteristics of “a 1:25,000 digital data set.”

2.3.6 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.

Raster representation of a field

In Figure 2.17, 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.

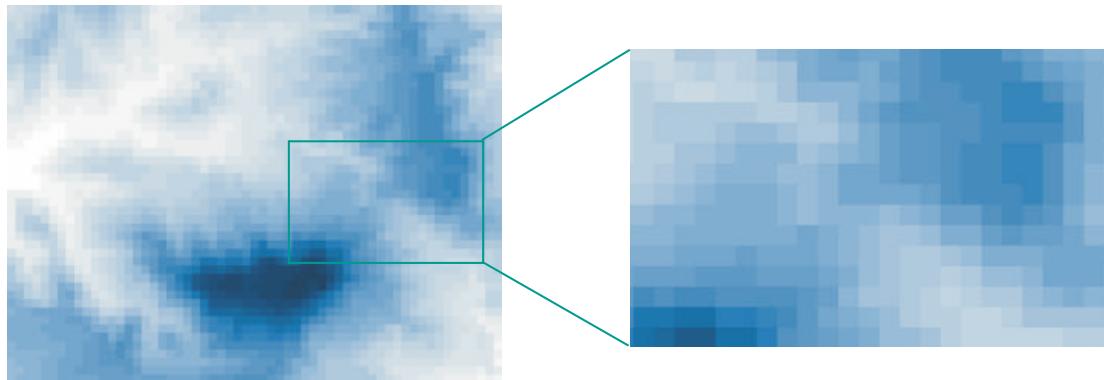


Figure 2.17: A raster representation (in part) of the elevation of the study area of Figure 2.2. Actual elevation values are indicated as shades of blue. The depicted area is the north-east flank of the mountain in the south-east of the study area. The right-hand side of the figure is a zoomed-in part 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 for interpreting 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 will assist in correctly representing peaks and mountain slope faces.

Vector representation of a field

We briefly mention a final representation for fields like elevation, but using a vector representation. 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 2.18. Both TINs and isoline representations use vectors.

Isoline



Figure 2.18: A vector-based elevation field representation for the study area of Figure 2.2. Indicated are elevation isolines at a resolution of 25 metres.

Data source: Department of Earth Systems Analysis (ESA, 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 representing this type of field is the better choice. Many GIS packages provide functions to generate an isoline visualization from a TIN.

2.3.7 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 2.2.4), and many of these parameters can be expressed in terms of vectors. However, tessellations are still commonly used for representing geographic objects as well, and we discuss why below.

Tessellations to represent geographic objects

Remotely sensed images are an important data source for GIS applications. Unprocessed digital images contain many 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 characterize 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 2.19 or urban land use classes as in the case of Figure 2.20. These figures illustrate the unprocessed images (a) as well as a classified version of the image (b).

Image classification

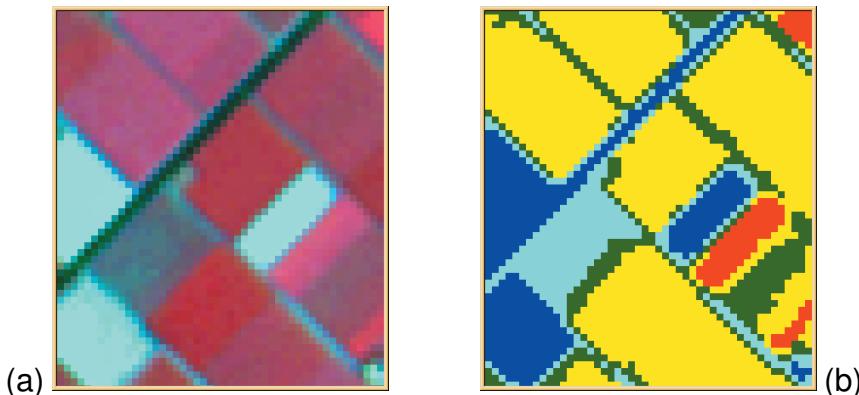


Figure 2.19: An unprocessed digital image (a) and a classified raster (b) of an agricultural area.

The application at hand may be interested only in geographic objects identified as potato fields (Figure 2.19(b), in yellow) or industrial complexes (Figure 2.20(b), in orange). 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.

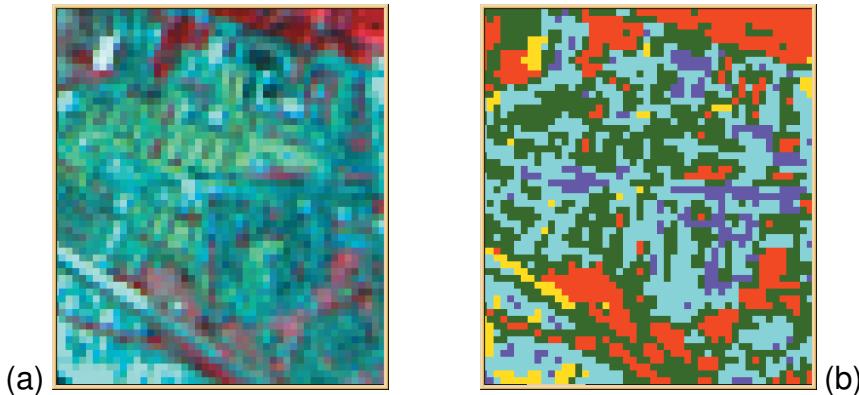


Figure 2.20: An unprocessed digital image (a) and a classified raster (b) of an urban area.

How the process of image classification takes place is not the subject of this book. It is dealt with in *Principles of Remote Sensing* [53]. 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 as jagged edges. 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 recognize these objects as points or lines.

Many GISs 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 2.21. Supported operations are connectivity operations and distance computations. There is again an issue of precision of such computations.

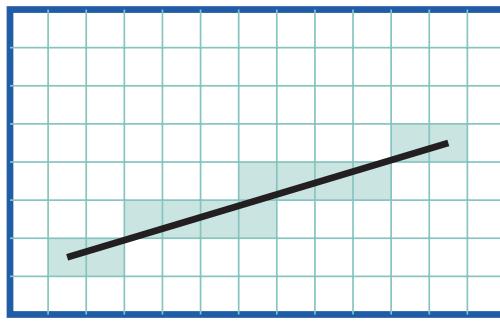


Figure 2.21: An actual straight line (in black) and its representation (light green cells) in a raster.

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 2.3.3, and a small example suffices at this stage.

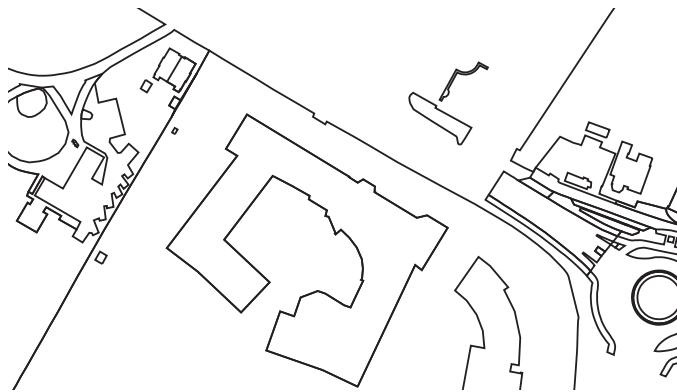


Figure 2.22: Various objects (buildings, bike and road lanes, railroad tracks) represented as area objects in a vector representation.

In Figure 2.22, 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.

2.4 Organizing and managing 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, we have purposefully avoided looking at how various sorts of spatial data are combined in a single system.

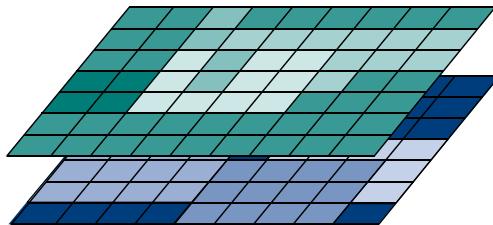


Figure 2.23: Different rasters can be overlaid to look for spatial correlations.

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. Usually, the data is organized so that similar elements are in a single data layer. For example, all telephone booth point objects would be in one layer, and all road line objects in another. A data layer contains spatial data—of any of the types discussed above—as well as attribute (or: thematic) data, which further describes the field or objects in the layer. Attribute data is quite often arranged in tabular form, maintained in some kind of geodatabase, as we will see in Chapter 3. An example of two field data layers is provided in Figure 2.23.

Management of attribute or thematic data

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 relationships* between different phenomena, requiring

computations which overlay one data layer with another. This is schematically depicted in Figure 2.24 for two different object layers. Field layers can also be involved in overlay operators. Chapter 3 will discuss the functions offered by GISs and database systems for data management in more detail.

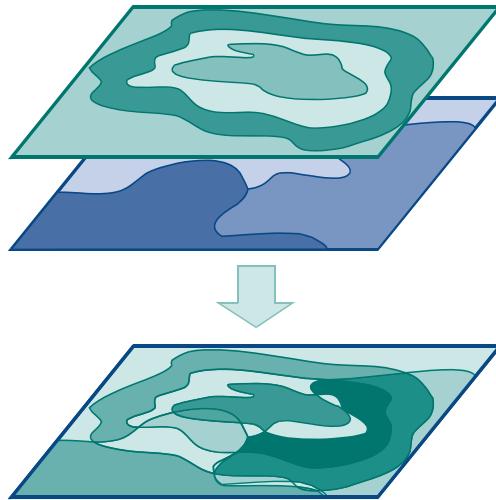


Figure 2.24: Two different object layers can be overlaid to look for spatial correlations, and the result can be used as a separate (object) layer.

2.5 The temporal dimension

Besides having geometric, thematic and topological properties, geographic phenomena are also *dynamic*; they change over time. For an increasing number of applications, these changes themselves are the key aspect of the phenomenon to study. Examples include identifying the owners of a land parcel in 1972, or how land cover in a certain area changed from native forest to pastures over a specific time period. We can note that some features or phenomena change slowly, such as geological features, or as in the example of land cover given above. Other phenomena change very rapidly, such as the movement of people or atmospheric conditions. For different applications, different scales of measurement will apply.

Dynamic phenomena

Examples of the kinds of questions involving time include:

- Where and when did something happen?
- How fast did this change occur?
- In which order did the changes happen?

The way we represent relevant components of the real world in our models can influence the kinds of questions we can or cannot answer. This chapter has already discussed representation issues for spatial features, but has so far ignored the problematic issues for incorporating time. The main reason lies in the fact that GISs still offer limited support for the representation of time. As a result, most studies require substantial efforts from the GIS user in data preparation and data manipulation. Also, besides representing an object or field in 2D or 3D

Time in GIS

space, the temporal dimension is of a continuous nature. Therefore in order to represent it in a computer, we have to ‘discretize’ the time dimension.

Spatiotemporal data models are ways of organizing representations of space *and* time in a GIS. Several representation techniques have been proposed in the literature. Perhaps the most common of these is a ‘snapshot’ state that represents a single point in time of an ongoing natural or man-made process. We may store a series of these snapshot states to represent change, but must be aware that this is by no means a comprehensive representation of that process. Further discussion of spatiotemporal data models is outside the scope of this book, and readers are referred to Langran [33] for a discussion of relevant concepts and issues. Here we will present a brief examination of different ‘concepts’ of time.

Representing time in GIS

- **Discrete and continuous time:** Time can be measured along a *discrete* or *continuous* scale. Discrete time is composed of discrete elements (seconds, minutes, hours, days, months, or years). 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’, and ‘after’.
- **Valid time and transaction time:** *Valid time* (or *world time*) is the time when an event really happened, or a string of events took place. *Transaction time* (or *database time*) is the time when the event was stored in the database or GIS. Observe that the time at which we store something in the database/GIS typically is (much) later than when the related event took place.
- **Linear, branching and cyclic time:** Time can be considered to be *linear*, ex-

tending from the past to the present ('now'), and into the future. This view gives a single time line. For some types of temporal analysis, *branching* time—in which different time lines from a certain point in time onwards are possible—and *cyclic* time—in which repeating cycles such as seasons or days of a week are recognized, make more sense and can be useful.

- **Time granularity:** When measuring time, we speak of *granularity* as the precision of a time value in a GIS or database (e.g. year, month, day, second, etc.). Different applications may obviously require different granularity. In cadastral applications, time granularity might well be a day, as the law requires deeds to be date-marked; in geological mapping applications, time granularity is more likely in the order of thousands or millions of years.
- **Absolute and relative time:** Time can be represented as *absolute* or *relative*. Absolute time marks a point on the time line where events happen (e.g. '6 July 1999 at 11:15 p.m.'). Relative time is indicated relative to other points in time (e.g. 'yesterday', 'last year', 'tomorrow', which are all relative to 'now', or 'two weeks later', which is relative to some other arbitrary point in time.).

Part of an example data set from a project investigating change is provided in Figure 2.25. The purpose of this particular study was to assess whether radar images are reliable resources for detecting the disappearance of primary forests [6]. This area of work is commonly known as *change detection*. Studies of this type are usually based on some 'model of change', which includes knowledge and hypotheses of how change occurs for the specific phenomena being studied. In this case, it included knowledge about speed of tree growth.

Change detection

In spatiotemporal analyses we consider changes of spatial and thematic attributes over time. We can keep the *spatial domain fixed* and look only at the attribute changes over time for a given location in space. We might be interested how land cover changed for a given location or how the land use changed for a given land parcel over time, provided its boundary did not change. On the other hand, we can keep the *attribute domain fixed* and consider the spatial changes over time for a given thematic attribute. In this case, we might want to identify locations that were covered by forest over a given period of time. 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*, a subject receiving increasing attention in the literature. Applications of moving object research include traffic control, mobile telephony, wildlife tracking, vector-borne disease control, and weather forecasting.

Spatiotemporal analysis

In these types of applications, the problem of *object identity* becomes apparent. When does a change or movement cause an object to disappear and become a new one? With wildlife this is quite obvious; with weather systems less so. But this should no longer surprise the reader: we have already seen that some geographic phenomena can be nicely described as objects, while others are better represented as fields.

Object identity

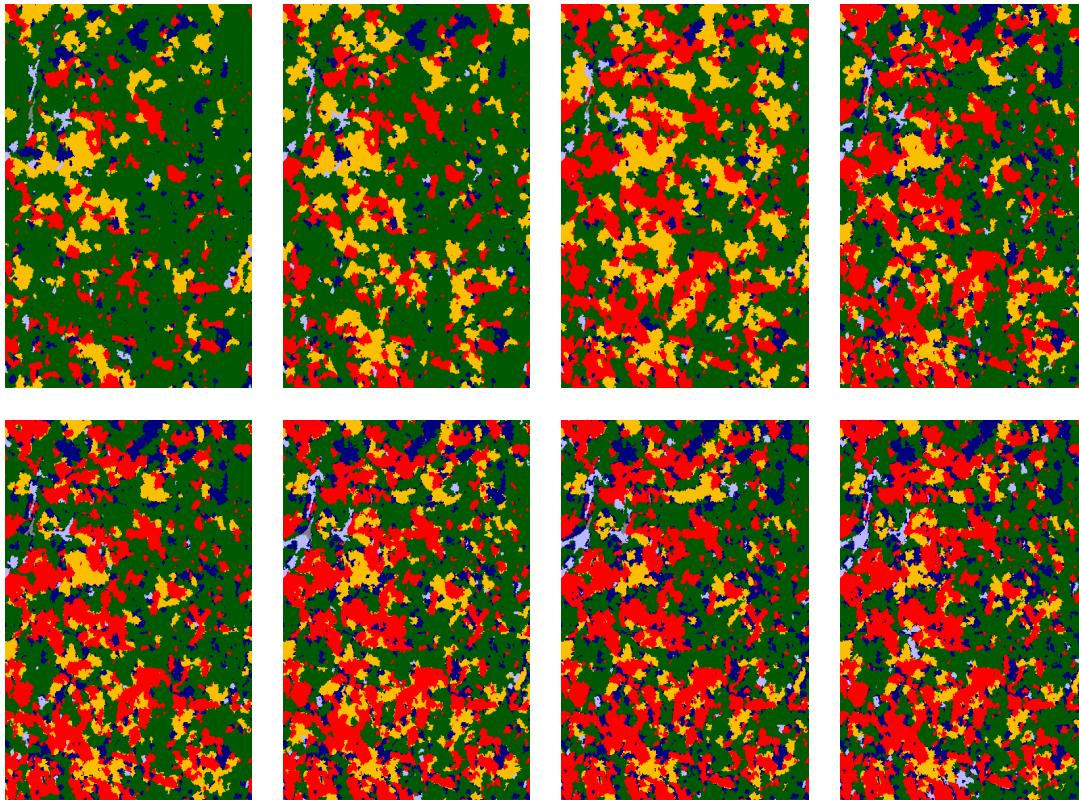


Figure 2.25: The change of land cover in a 9×14 km study site near San José del Guaviare, div. Guaviare, Colombia, during a study conducted in 1992–1994 by Bijker [6]. A time series of ERS-1 radar images after application of (1) image segmentation, (2) rule-based image classification, and (3) further 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.

Summary

Geographic phenomena are present in the real world that we study; their computer representations only live inside computer systems. This chapter has discussed different types of geographic phenomena, and examined the ways that these can be represented in a computer system, such as a GIS.

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 include location, shape, size and orientation. The dimension of an object is a fundamental part of its shape parameter: is it a point, line, area or volume object? In all cases, a representation of the boundary of the object (whether crisp or fuzzy), is often used in GIS.

The second half of the chapter 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.
2. Take the list of question 1 and identify which phenomena are fields and which are objects. Which of your example objects are crisp?
3. There is 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.
4. 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.
5. On page 70, 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 a 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.)
6. On page 108, 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*.

7. What colour is the northwest quadrant of the outermost northeast quadrant of Figure 2.7? 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?
8. Make an educated guess at the elevation of location *P* in Figure 2.8. 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 2.9 (first the left one, then the right one).
9. Explain how many line objects and how many line segments are illustrated in Figure 2.12. Complete the table on the left, using a numbering of vertices that you have made up yourself for Figure 2.11.
10. In this 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?
11. In Figure 2.21, 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 2.21? What can be said in general about the two lengths?



12. 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?
13. What observations can be made from a visual interpretation of Figure 2.25? What changes do you ‘detect’? Which stages of change?



Chapter 3

Data management and processing systems

The ability to manage and process spatial data is a critical component for any functioning GIS. Simply put, data processing systems refer to hardware and software components which are able to process, store and transfer data. This chapter discusses the components of systems that facilitate the management and processing of geoinformation. In order to provide a brief background to the discussion, the chapter begins with a brief discussion of computer hardware and software trends.

In Section 3.4, we discuss database management systems (DBMSs), and illustrate some principles and methods of data extraction from a database. The final section of the chapter (Section 3.5) looks at the merging of GIS and DBMS, and

the emergence of spatial databases in recent years. It notes their key advantages, and briefly illustrates the use of a spatial database for data storage and processing.

3.1 Hardware and software trends

Advances in computer hardware seem to take place at an ever-increasing rate. Every several months, a faster, more powerful processor generation replaces the previous one. Computers are also becoming increasingly portable, while offering this increased performance. The computing power that we have available in today's handheld computers is a multiple of the performance that the first PC had when it was introduced in the early 1980's. In fact, current PCs have orders of magnitude more memory and storage capacity than the so-called minicomputers of 25 years ago. To illustrate this trend: compare a typical early 1980's PC with a 2 MHz CPU, 128 kbytes of main memory, and a 10 MByte hard disk to the current generation of desktop PC's. Table 3.1 shows the list of standard unit prefixes for reference purposes.

Handheld PC's

Computers are also becoming increasingly affordable. Hand-held computers are now commonplace in business and personal use, equipping field surveyors with powerful tools, complete with GPS capabilities for instantaneous georeferencing. To support these hardware trends, software providers continue to produce application programs and operating systems that, while providing a lot more functionality, also consume significantly more memory. In general, software technology has developed somewhat slower and often cannot fully utilise the possibilities offered by the exponentially growing hardware capabilities. Existing software obviously performs better when run on faster computers.

Alongside these trends, there have also been significant developments in computer networks. In essence, today almost any computer on Earth can connect to some network, and contact computers virtually anywhere else, allowing fast and reliable exchange of (spatial) data. Mobile phones are more and more frequently

being used to connect to computers on the Internet. The UMTS protocol (Universal Mobile Telecommunications System), allows digital communication of text, audio, and video at a rate of approximately 2 Mbps. The new HSDPA protocol offers up to 10 times this speed. Looking at these developments it is clear that the combination of a GPS receiver, a portable computer and mobile phone has already dramatically changed our world, certainly so for out-of-office activities of Earth science professionals .

Mobile communication

<i>prefix</i>	m	c	d	h	k	M	G	T	P	E
<i>name</i>	milli	centi	deci	hecto	kilo	mega	giga	tera	peta	exa
<i>factor</i>	10^{-3}	10^{-2}	10^{-1}	10^2	10^3	10^6	10^9	10^{12}	10^{15}	10^{18}

Table 3.1: Commonly used unit prefixes

Bluetooth version 2.0 is a standard that offers up to 3 Mbps connections, especially between palm- and laptop computers and their peripheral devices, such as a mobile phone, GPS or printer at short range. Wireless LANs (Local Area Networks), under the so-called WiFi standard, nowadays offer a bandwidth of up to 108 Mbps on a single connection point, *to be shared* between computers. They are more and more used for constructing a computer network in office buildings and in private homes.

Wireless LAN and WiFi

When the medium of communication is not the air, but copper or fibre optics cables (structured networks), the picture is a different one. Standard 'Dial-up' telephone modems allow rates up to 56 kbps. Digital telephone links (ISDN) support much higher rates: up to 1.5 Mbps. ADSL technology widely available through telephone companies on standard copper-wire networks supports transfer rates anywhere between 2 and 20 Mbps towards the customer (downstream), and between 1 and 8 Mbps towards the network (upstream) depending on the internet provider and quality of the network infrastructure.

Structured networks

Wide-area computer networks (national, continental, global) have a capacity of several Gbps. ITC's dedicated Local Area Network (LAN), which is partially fibre optics-based, supports a transmission rate locally of 1 Gbps. Since fibre optic cables in principle support rates of various Gbps, it is unlikely that this bandwidth capacity will be exceeded in the very near future.

3.2 Geographic information systems

It was identified in Chapter 1 that a GIS provides a range of capabilities to handle georeferenced data, including:

1. Data capture and preparation,
2. Data management (storage and maintenance),
3. Data manipulation and analysis, and
4. Data presentation.

For many years, analogue data sources were used, processing was done manually, and paper maps were produced. The introduction of modern techniques 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 centered around geographic information systems.

Typical planning projects require data sources, both spatial and non-spatial, from different national institutes, like national mapping agencies, geological, soil, and forest survey institutes, and national census bureaus. The data sources obtained may be from different time periods, and the spatial data may be in different scales or projections. With the help of a GIS, the spatial data can be stored in digital form in world coordinates. This makes scale transformations unnecessary, and the conversion between map projections can be done easily with the software. With the spatial data thus prepared, spatial analysis functions of the GIS can then be applied to perform the planning tasks.

Data requirements

What remains is to pay careful attention to the quality (or lack of it) in the different datasets, to ensure that unnecessary error is not being introduced. Chapter 5 discusses these issues in more detail. In this chapter we will focus on GIS software and ways to manage spatial and attribute data.

3.2.1 GIS software

As noted previously, GIS can be considered to be a data store (i.e. a system that stores spatial data), a toolbox, a technology, an information source or a field of science. 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.

The use of tools for problem solving is one thing, but the production of these tools is something quite different. Not all tools are equally well-suited for a particular application, and they can be improved and perfected to better serve a particular need or application. The discipline of *geographic information science* is driven by the use of our GIS tools, and these are in turn improved by new insights and information gained through their application in various scientific fields. *Spatial information theory* is one such field, which focuses specifically on providing the background for the production of tools for the handling of spatial data.

GIS tools

All GIS packages available on the market have their strengths and weaknesses, typically resulting from the development history and/or intended application domain(s) of the package. Some GISs have traditionally focused more on support for raster-based functionality, others more on (vector-based) spatial objects. We can safely state that any package that provides support for only rasters or only objects, is *not* a complete GIS. Well-known, full-fledged GIS packages include ILWIS, Intergraph's GeoMedia, ESRI's ArcGIS, and MapInfo from Map-Info Corp. Several of these systems are used within ITC in practical sessions of the Principles of GIS teaching module. This textbook attempts to describe the field of GIS independently from specific software packages, as 'principles'

Industry-standard GIS
packages

should be useful to users of any package.

There is no particular GIS package which is necessarily 'better' than another one: this depends on factors such as the intended application, and the expertise of its user. ILWIS's traditional strengths are in raster processing and scientific spatial data analysis, especially in project-based GIS applications. Intergraph, ESRI and MapInfo products have been known better for their 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, fails to do justice to any of these packages, and it is only after extended use that their strengths, and sometimes weaknesses, might become clear.

3.2.2 GIS architecture and functionality

We have already noted that a geographic information system in the wider sense consists of software, data, people, and an organization in which it is used. Before moving on, we should also note that organizational factors will define the context and rules for the capture, processing and sharing of geoinformation, as well as the role which GIS plays in the organization as a whole. In the remainder of this chapter we focus on the architecture and functional components of GIS software.

Role of GIS in organizations

As noted above, a GIS consists of several *functional components*—components which support key GIS functions. These are data capture and preparation, data storage, data analysis, and presentation of spatial data. Figure 3.1 shows a diagram of these components, with arrows indicating the data flow in the system. For a particular GIS, each of these components may provide many or only a few functions. Arguably, the system should not be called a geographic information system if any one of these components is missing. It is important to note however, that the same function may be offered by different components of the GIS: for instance, data capture and data storage may have functions in common, and the same holds for data preparation and data analysis.

The following sections briefly describe these components, focussing on storage and maintenance. Later in this chapter, we will discuss the role of DBMS and more specifically, spatial databases, in the storage and maintenance of geospatial data. A more detailed treatment of the other functional components in Figure 3.1 can be found in follow-up chapters.

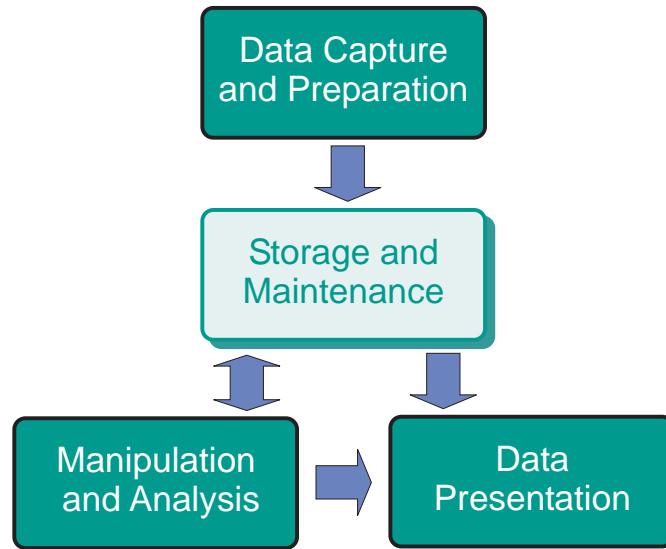


Figure 3.1: Functional components of a GIS

3.2.3 Spatial Data Infrastructure (SDI)

For reasons that include efficiency and legislation, many organizations are forced to work in a cooperative setting in which geographic information is obtained from, and provided to, partner organizations and the general public. The sharing of spatial data between the various GISs in those organizations is of key importance and aspects of data dissemination, security, copyright and pricing require special attention. The design and maintenance of a Spatial Data Infrastructure (SDI) deals with these issues.

Data sharing

In [42] an SDI is defined as “the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data”. Fundamental to those arrangements are—in a wider sense—the agreements between organizations and in the narrow sense, the agreements between software systems on *how* to share the geographic information. In SDI, standards are often the starting point for those agreements. Standards exist for all facets of GIS, ranging from data capture to data presentation. They are developed by different organizations, of which the most prominent are the International Organization for Standardisation (ISO) and the Open Geospatial Consortium (OGC).

Standards

Typically, an SDI provides its users with different facilities for finding, viewing, downloading and processing data. Because the organizations in an SDI are normally widely distributed over space, computer networks are used as the means of communication. With the development of the internet, the functional components of GIS have been gradually become available as web-based applications. Much of the functionality is provided by so called geo-webservices, software programs that act as an intermediate between geographic data(bases) and the

Geo-webservices

users of the web. Geo-webservices can vary from a simple map display service to a service which involves complex spatial calculations. For their spatial data handling, these services commonly use standardized raster and vector representations following the abovementioned standards.

3.3 Stages of spatial data handling

3.3.1 Spatial data capture and preparation

The functions for capturing data 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 spatial data sets are derived. Surveys of the study area often need to be conducted for data that cannot be obtained with remote sensing techniques, or to validate data thus obtained.

Traditional techniques for obtaining spatial data, typically from paper sources, included *manual digitizing* and *scanning*. Table 3.2 lists the main methods and devices used for data capture. In recent years there has been a significant increase in the availability and sharing of digital (geospatial) data. As discussed above, various media and computer networks play an important role in the dissemination of this data, particularly the internet.

Digitizing and scanning

The data, once obtained in some digital format, may not be quite ready for use in the system. This may be because the format obtained from the capturing process is not quite the format required for storage and further use, which means that some type of data conversion is required. In part, this problem may also arise when the captured data represents only raw base data, out of which the real data objects of interest to the system still need to be constructed. For example, semi-automatic digitizing may produce line segments, while the application's requirements are that non-overlapping polygons are needed. A build-and-verification phase would then be needed to obtain these from the captured lines.

Data conversion

Issues related to data acquisition and preparation are discussed in greater detail

<i>Method</i>	<i>Devices</i>
Manual digitizing	<ul style="list-style-type: none">• coordinate entry via keyboard• digitizing tablet with cursor• mouse cursor on the computer monitor (heads-up digitizing)• (digital) photogrammetry
Automatic digitizing	<ul style="list-style-type: none">• scanner
Semi-automatic digitizing	<ul style="list-style-type: none">• line-following software
Input of available digital data	<ul style="list-style-type: none">• CD-ROM or DVD-ROM• via computer network or internet (including geo-webservices)

Table 3.2: Spatial data input methods and devices used

in Chapter 5.

3.3.2 Spatial data storage and maintenance

The way that data is stored plays a central role in the processing and the eventual understanding of that data. In most of the available systems, spatial data is organized in layers by theme and/or scale. For instance, the data may be organized in thematic categories, such as land use, topography and administrative subdivisions, or according to map 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 naturally as possible. In a GIS, features are represented with their (geometric and non-geometric) attributes and relationships. The geometry of features is represented with primitives of the respective dimension: a windmill probably as a point, an agricultural field as a polygon. The primitives follow either the vector, as in the example, 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) cells, mostly as a square tessellation of dimension two or three. These cells are called either *cells* or *pixels* in 2D, and *voxels* in 3D. The data indicates for every cell which real world feature it covers, in case it represents a discrete field. In case of a continuous field, the cell holds a representative value for that field. Table 3.3 lists advantages and disadvantages of raster and vector representations.

Data organization

Cells, pixels and voxels

The storage of a raster is, in principle, straightforward. It is stored in a file as a long list of values, one for each cell, preceded by a small list of extra data (the so-called ‘file header’) that informs how to interpret the long list. The order of the cell values in the list can be—but need not be—left-to-right, top-to-bottom.

Raster representation	Vector representation
<i>advantages</i>	
<ul style="list-style-type: none"> • simple data structure • simple implementation of overlays • efficient for image processing 	<ul style="list-style-type: none"> • efficient representation of topology • adapts well to scale changes • allows representing networks • allows easy association with attribute data
<i>disadvantages</i>	
<ul style="list-style-type: none"> • less compact data structure • difficulties in representing topology • cell boundaries independent of feature boundaries 	<ul style="list-style-type: none"> • complex data structure • overlay more difficult to implement • inefficient for image processing • more update-intensive

Table 3.3: Raster and vector representations compared

This simple encoding scheme is known as *row ordering*. The header of the raster file will typically inform how many rows and columns the raster has, which encoding scheme is used, and what sort of values are stored for each cell. Raster files can be quite big data sets. For computational reasons, it is wise to organize the long list of cell values in such a way that spatially nearby cells are also near to each other in the list. This is why other encoding schemes have been devised. The reader is referred to [34] for a more detailed discussion.

Raster encoding

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 2.12, where a boundary

model for polygon objects was illustrated. Similar structures are in use for line objects. For further, advanced reading, please see [49].

GIS software packages provide support for both spatial and attribute data, i.e. they accommodate spatial data storage using a vector approach, and attribute data using tables. Historically, however, *database management systems* (DBMSs) have been based on the notion of tables for data storage. For some time, substantial GIS applications have been able to link to an external database to store attribute data and make use of its superior data management functions. Currently, All major GIS packages provide facilities to link with a DBMS and exchange attribute data with it. Spatial (vector) and attribute data are still sometimes stored in separate structures, although they can now be stored directly in a spatial database. More detail on these issues is provided in Section 3.5.

DBMS and spatial databases

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 to have an up-to-date stored data set available. After a major earthquake, for instance, we may have to update our road network data to reflect that roads have been washed away, or have otherwise become impassable.

Data maintenance

The need for updating spatial data stems from the requirements that the data users impose, as well as the fact that many aspects of the real world change continuously. These data updates can take different forms. It may be that a complete, new survey has been carried out, from which an entirely new data set is derived that will replace the current set. Such a situation is typical if the spatial data originates from remotely sensed data, for example, a new vegetation

cover set, or a new digital elevation model. It may also be that local (ground) surveys have revealed local changes, for instance, new constructions, or changes in land use or ownership. In such cases, local change to the large spatial data set is more typically required. Such local changes should respect matters of data consistency, i.e. they should leave other spatial data within the same layer intact and correct.

3.3.3 Spatial query and analysis

The most distinguishing parts of a GIS are its functions for spatial analysis, i.e. operators that use spatial data to derive new geoinformation. *Spatial queries* and *process models* play an important role in this functionality. One of the key uses of GISs has been to support spatial decisions. *Spatial decision support systems* (SDSS) are a category of information systems composed of a database, GIS software, models, and a so-called knowledge engine which allow users to deal specifically with locational problems.

SDSS

In a GIS, data are usually grouped into 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 provide answers to user questions. GIS functions are used for maintenance of the data, and for analysing the data in order to infer information from it. *Analysis* of spatial data can be defined as computing new information that provides new insight from the existing, stored spatial data.

Spatial data analysis

Consider an example from the domain of road construction. In mountainous areas this is a complex engineering task with many cost factors, which include 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. Maintenance and analysis of attribute data is discussed further in Section 3.4.

The exact nature of the analysis will depend on the application requirements, but computations and analytical functions operate on both spatial and non-spatial data. Chapter 6 discusses these issues in more detail. For now, we will focus on

the last stage of Figure 3.1.

3.3.4 Spatial data presentation

The presentation of spatial data, whether in print or on-screen, in maps or in tabular displays, or as 'raw data', is closely related to the disciplines of cartography, printing and publishing. The presentation may either be an end-product, for example as a printed atlas, or an intermediate product, as in spatial data made available through the internet.

<i>Method</i>	<i>Devices</i>
Hard copy	<ul style="list-style-type: none">• printer• plotter (pen plotter, ink-jet printer, thermal transfer printer, electrostatic plotter)• film writer
Soft copy	<ul style="list-style-type: none">• computer screen
Output of digital data sets	<ul style="list-style-type: none">• magnetic tape• CD-ROM or DVD• the Internet

Table 3.4: Spatial data presentation

Table 3.4 lists several different methods and devices used for the presentation of spatial data. Cartography and scientific visualization make use of these methods and devices to produce their products. Chapter 7 is devoted to visualization techniques for spatial data.

3.4 Database management systems

A *database* is a large, computerized collection of structured data.

In the non-spatial domain, databases have been in use since the 1960's, for various purposes like bank account administration, stock monitoring, salary administration, order bookkeeping, and flight reservation systems to name just a few. The common denominator between these applications is that the amount of data is usually quite large, but the data itself has a simple and regular structure.

Designing a database is not an easy task. Firstly, one has to consider carefully what the database purpose is, and who its users will be. Secondly, 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*. Lastly, data can be entered into the database. It is important to keep the data up-to-date, and it is therefore wise to set up the processes for this, and make someone responsible for regular maintenance of the database. Documentation of the database design and set-up is crucial for an extended database life. Many enterprise databases tend to outlive the professional careers of their original designers.

Database design and
maintenance

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 the set-up of 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 offered by DBMSs. Many standard PCs

are equipped with a DBMS called MS Access. This package offers a useful set of functions, and the capacity to store terabytes of information.

3.4.1 Reasons for using a DBMS

There are various reasons why one would want to use a DBMS for data storage and processing.

- A DBMS supports the storage and manipulation of *very large data sets*.

Some data sets are so big that storing them in text files or spreadsheet files becomes too awkward for use in practice. The result may be that finding simple facts takes minutes, and performing simple calculations perhaps even hours. A DBMS is specifically designed for this purpose.

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

For instance, an important aspect of data correctness is data entry checking: ensuring that the data that is entered into the database does not contain obvious errors. For instance, since we know the study area we are working in, we also know the range of possible geographic coordinates, so we can ensure the DBMS checks 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 design of a database.

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

Large data sets are built up over time, which means that substantial investments are required to create and maintain 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.

Moreover, for different users of the database, different views on 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. They may all be using the database at the same time, without affecting each other's activities. This DBMS function is called *concurrency control*.

- A DBMS provides a high-level, *declarative query language*.¹

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.

- 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.

¹The 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-friendly because the user need not care about the 'how' and can focus on the 'what'.

The most prominent data model is the *relational data model*. We discuss it in full in Section 3.4.3. 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 all times.

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 the control of *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*—can lead to situations in which stored facts may contradict each other, causing reduced usefulness of the data. Redundancy, however, is not necessarily always problematic, as long as we specify where it occurs so that it can be controlled for.

3.4.2 Alternatives for data management

The decision whether or not to use a DBMS will depend, among other things, on how much data there is or will be, what type of use will be made of it, and how many users might be involved.

On the small-scale side of the spectrum—when the data set is small, its use 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 small set of simple field observations. Text files offer no support for data analysis whatsoever, except perhaps in alphabetical sorting.

If our data set is still small and numeric by nature, and we have a single type of use in mind, a spreadsheet program will suffice. This might be the case if we have a number of field observations with measurements that we want to prepare for statistical analysis, for example. 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. It should also be noted that spreadsheets do not accommodate concurrent use of the data set well, although they do support some data analysis, especially when it comes to calculations over a single table, like averages, sums, minimum and maximum values.

All such computations are usually 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, some expertise and significant amounts of time are usually required to make this happen.

3.4.3 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 used to define the database are *attributes*, *tuples* and *relations*. Computer programs either perform data extraction from the database without altering it, in which case we call them *queries*, or they change the database contents, and we speak of *updates* or *transactions*. The technical terms surrounding database technology are defined below.

Let us look at a tiny database example from a cadastral setting. It is illustrated in Figure 3.2. This database consists of three tables, one for storing people's 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).

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 3.2: 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 3.2, 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*.

The example relations provided in the figure should clarify this. The Private-Person table has three tuples; the Surname attribute value for the first tuple illustrated is 'Garcia.'

The phrase 'that are similarly shaped' takes this a little bit further. It requires that all values for the same attribute come from a single domain of values. An attribute's *domain* is a (possibly infinite) set of atomic values such as the set of integer number values, the set of real number values, etc. 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.

Attribute domain

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

Table 3.5: The relation schemas for the three tables of the database in Figure 3.2.

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
3. Set the *domain* of each attribute.

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 3.2, the relation schemas are given in Table 3.5. Underlined attributes (and their domains) indicate the *primary key* of the relation, which will be defined and discussed below. Relation schemas are stable, and will rarely change over time. This is not true of the tuples stored in tables: they, typically, are often changing, either because new tuples are added, others are removed, or yet others will see changes in their attribute values.

Primary key

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: It is possible to count how many tuples there are. Figure 3.2 gives us a single *database instance*, i.e. one relation

Relation instance

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

We have already stated that database systems are particularly good at storing large quantities of data. (Note: our example database is not even small, it is tiny!) The DBMS must support quick searches amongst many tuples. This is why the relational data model uses the notion of a 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 no more than one tuple in the table with that combination of values. It remains possible that there is no tuple for the given combination. In our example database, the set {TaxId, Surname} is a key of the relation **PrivatePerson**: if we know both a TaxId and a Surname value, we will find at most one tuple with that combination of values.

Every relation has a key, though possibly it is the combination of all attributes. Such a large key, 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.²

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

²As an aside, note that an attribute such as **AreaSize** in relation **Parcel** is *not* a key, although it appears to be one in Figure 3.2. The reason is that some day there could be a second parcel with size 435, giving us two parcels with that value.

The diagram illustrates a database schema with two relations: **TitleDeed** and **Parcel**. A dashed arrow points from the **TitleDeed** table to the **Parcel** table, indicating a relationship between them. The **TitleDeed** table has columns **TitleDeed**, **Plot**, **Owner**, and **DeedDate**. The **Parcel** table has columns **Parcel**, **Pid**, **Location**, and **AreaSize**. Two tuples in the **TitleDeed** table have their **Plot** attribute values circled in red: 2109 and 3421. These circled values correspond to the **Pid** values of two tuples in the **Parcel** table: 2109 and 3421 respectively.

TitleDeed	Plot	Owner	DeedDate
2109	101-367	18/12/1996	
8871	101-490	10/01/1984	
1515	134-788	01/09/1991	
3421	101-367	25/09/1996	

Parcel	Pid	Location	AreaSize
3421	2001	435	
8871	1462	550	
2109	2323	1040	
1515	2003	245	

Figure 3.3: 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). 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 3.3. 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! A foreign key must have as many attributes as the primary key that it refers to.

Foreign key

3.4.4 Querying a relational database

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

The three query operators have some traits 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 3.4(a).

Tuple selection works like a filter: it allows tuples that meet the selection condition to pass, and disallows tuples that do not meet the condition.

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.

Selection condition

A second operator is also illustrated in Figure 3.4. 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.

Attribute projection works like a tuple formatter: it passes through all tuples of the input, but reshapes each of them in the same way.

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 3.4 are written in this language as follows:

```
SELECT *  
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.

```
SELECT PId, Location  
FROM   Parcel
```

(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.

SQL

Queries like the two above do *not* create stored tables in the database. This is why the result tables have no name: they are virtual tables. The result of a query is a table that is shown to the user who executed the query. Whenever the user closes her/his view on the query result, that result is lost. The SQL code for the query is stored, however, for future use. The user can re-execute the query again to obtain a view on the result once more.

Virtual tables

Our third query operator differs from the two above in that it requires two input relations. The operator is called the *join*, and is illustrated in Figure 3.5.

The join operator takes two input relations and produces one output relation, gluing two tuples together (one from each input relation), to form a bigger tuple, if they meet a specified condition.

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 to a tuple from the second input relation. The join operator uses a condition that expresses which tuples from the first relation are combined (‘glued’) with which tuples from the second. The example of Figure 3.5 combines TitleDeed tuples with Parcel tuples, but only those for which the foreign key Plot matches with primary key PId.

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 operator for extracting sensible information from a database. The strength of the above operators hides in the fact that they can be combined to produce more advanced and useful query definitions. We provide a final example to illustrate this. Take another look at the join of Figure 3.5. 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.

Join condition

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 3.6.

Finally, we may look at the SQL statement that would give us the query of Figure 3.6. It can be written as

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

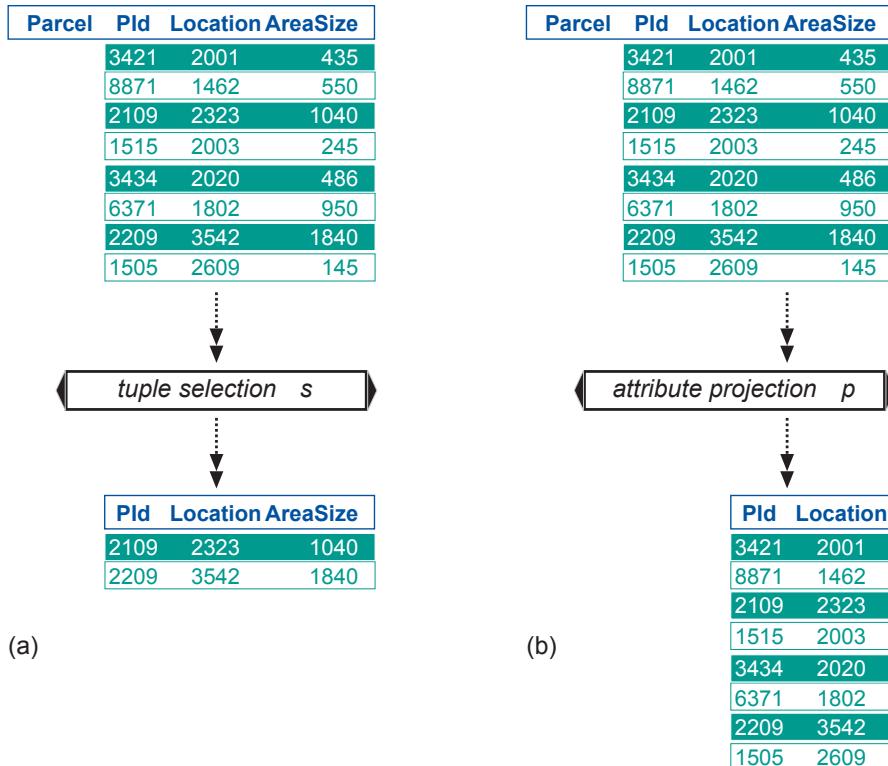


Figure 3.4: The two unary query operators: (a) tuple selection has a single table as input and produces another table with less tuples. Here, the condition was that AreaSize must be over 1000; (b) attribute projection has a single table as input and produces another table with fewer attributes. Here, the projection is onto the attributes PId and Location.

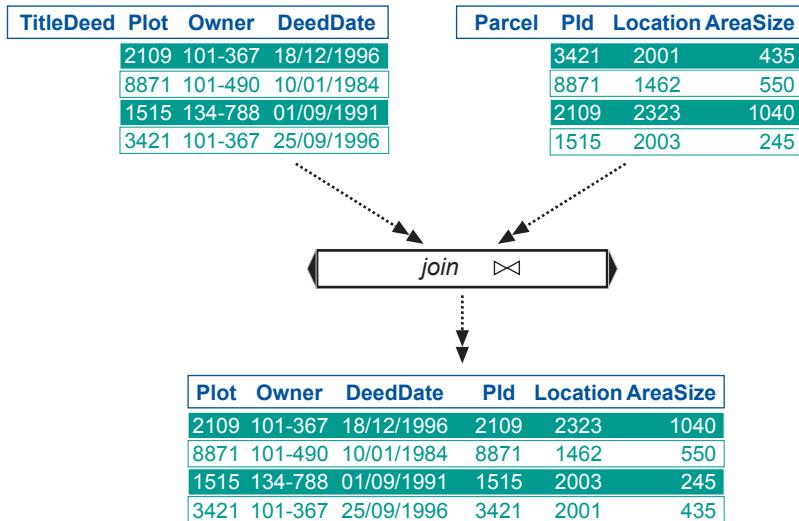


Figure 3.5: The essential binary query operator: join. The join condition for this example is *TitleDeed.Plot=Parcel.Pid*, which expresses a foreign key/key link between *TitleDeed* and *Parcel*. The result relation has $3 + 3 = 6$ attributes.

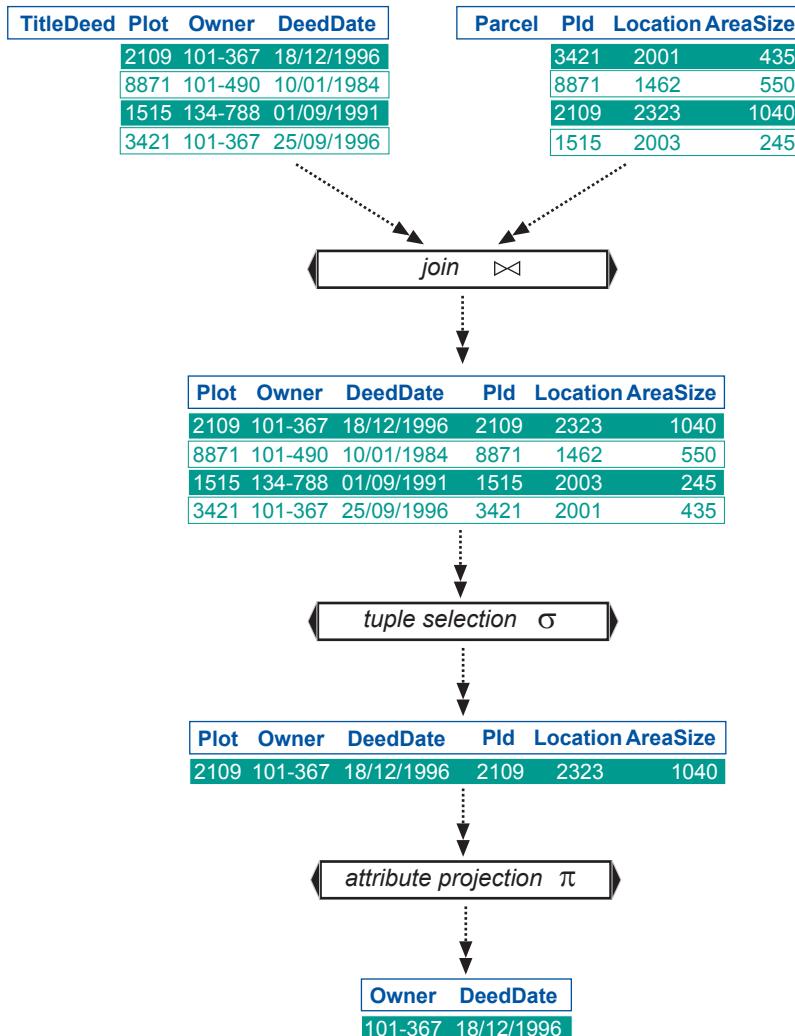


Figure 3.6: A combined selection/projection/join query, selecting owners and deed dates for parcels with a size larger than 1000. The join is carried out first, then follows a tuple selection on the result tuples of the join. Finally, an attribute projection is carried out.

3.5 GIS and spatial databases

3.5.1 Linking GIS and DBMS

GIS software provides support for *spatial* data and thematic or *attribute* data. GISs have traditionally stored spatial data and attribute data separately. This required the GIS to provide a link between the spatial data (represented with rasters or vectors), and their non-spatial attribute data. The strength of GIS technology lies in its built-in ‘understanding’ of geographic space and all functions that derive from this, for purposes such as storage, analysis, and map production. GIS packages themselves can store tabular data, however, they do not always provide a full-fledged query language to operate on the tables.

Storing spatial and attribute data

DBMSs have a long tradition in handling attribute (i.e. administrative, non-spatial, tabular, thematic) data in a secure way, for multiple users at the same time. Arguably, DBMSs offer much better table functionality, since they are specifically designed for this purpose. A lot of the data in GIS applications is attribute data, so it made sense to use a DBMS for it. For this reason, many GIS applications have made use of external DBMSs for data support. In this role, the DBMS serves as a centralized data repository for all users, while each user runs her/his own GIS software that obtains its data from the DBMS. This meant that a GIS had to link the spatial data represented with rasters or vectors, and the attribute data stored in an external DBMS.

External DBMS

With raster representations, each raster cell stores a characteristic value. This value can be used to look up attribute data in an accompanying database table. For instance, the land use raster of Figure 3.7 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 similarity with the key/foreign key concept in relational databases.

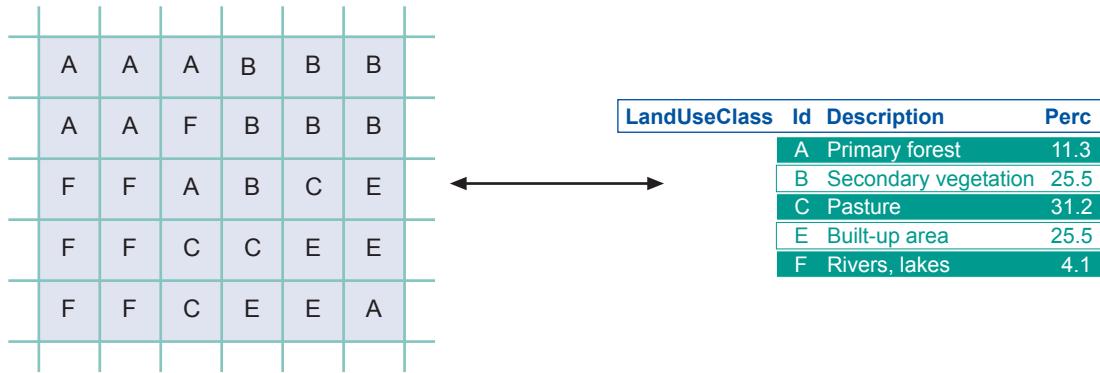


Figure 3.7: A raster representing land use and a related table providing full text descriptions (amongst others) of each land use class.

With vector representations, our spatial objects—whether they are points, lines or polygons—are automatically given a unique identifier by the system. This identifier is usually just called the object ID or feature ID and is used to link the spatial object (as represented in vectors) with its attribute data in an attribute table. The principle applied here is similar to that in raster settings, but in this case each object has its own identifier. The ID in the vector system functions as a key, and any reference to an ID value in the attribute database is a foreign key reference to the vector system. For example, in Figure 3.8, parcel is a table with attributes, linked to the spatial objects stored in a GIS by the Location column. Obviously, several tables may make references to the vector system, but it is not uncommon to have some main table for which the ID is actually also the key.

Linking objects and tables

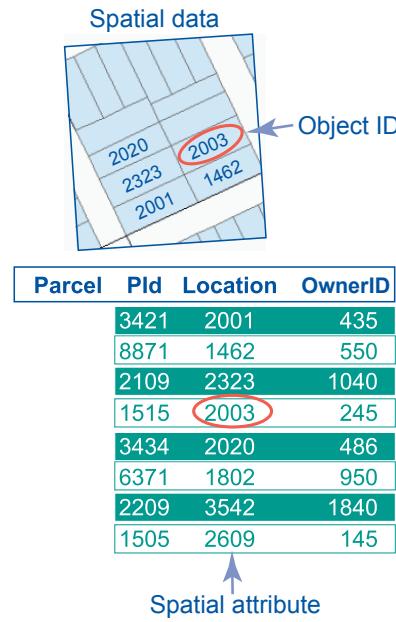


Figure 3.8: Storage and linking of vector attribute data between a GIS and DBMS.

3.5.2 Spatial database functionality

DBMS vendors have over the last 20 years recognized the need for storing more complex data, like spatial data. The main problem was that there is additional functionality needed by DBMS in order to process and manage spatial data. As the capabilities of our hardware to process information has increased, so too has the desire for better ways to represent and manage spatial data. During the 1990's, *object-oriented* and *object-relational* data models were developed for just this purpose. These extend standard relational models with support for objects, including 'spatial' objects.

Currently, GIS software packages are able to store spatial data using a range of commercial and open source DBMSs such as Oracle, Informix, IBM DB2, Sybase, and PostgreSQL, with the help of spatial extensions. Some GIS software have integrated database 'engines', and therefore do not need these extensions. ESRI's ArcGIS, for example, has the main components of the MS Access database software built-in. This means that the designer of a GIS application can choose whether to store the application data in the GIS or in the DBMS. Spatial databases, also known as *geodatabases*,³ are implemented directly on existing DBMSs, using extension software to allow them to handle spatial objects.

Spatial DBMS

A *spatial database* allows users to store, query and manipulate collections of spatial data.

There are several advantages in doing this, as we will see below. Put simply,

³Often, the term 'geodatabase' is used to refer to a specific kind of spatial database created with ESRI's ArcGIS software. Here we use it to refer to spatially-enabled DBMS in general.

spatial data can be stored in a special database column, known as the geometry column, (or feature or shape, depending on the specific software package), as shown in Figure 3.9. This means GISs can rely fully on DBMS support for spatial data, making use of a DBMS for data query and storage (and multi-user support), and GIS for spatial functionality. Small-scale GIS applications may not require a multi-user capability, and can be supported by spatial data support from a personal database.

Parcel	PId	Geometry	OwnerID
3421		"MULTIPOLYGON(((257462.704979333 464780.750851061,257463.89798...)))"	435
8871		"MULTIPOLYGON(((257409.813950544 464789.91585049,257407.896903...)))"	550
2109		"MULTIPOLYGON(((257785.714911912 464796.839972167,257782.59794...)))"	1040
1515		"MULTIPOLYGON(((257790.672100448 464807.13792585,257788.608078...)))"	245
3434		"MULTIPOLYGON(((257435.527950478 464803.92887633,257428.254887...)))"	486
6371		MULTIPOLYGON(((257432.476077854 464813.848852072,257433.147910...)))	950
2209		"MULTIPOLYGON(((257444.888027332 464826.555046319,257446.43201...)))"	1840
1505		"MULTIPOLYGON(((256293.760107491 464935.203846095,256292.00881...)))"	145

Figure 3.9: Geometry data stored directly in a spatial database table.

A geodatabase allows a wide variety of users to access large data sets (both geographic and alphanumeric), and the management of their relations, guaranteeing their integrity. The Open Geospatial Consortium (OGC) has released a series of standards relating to geodatabases that (amongst other things), define:

- Which tables must be present in a spatial database (i.e. geometry columns table and spatial reference system table)
- The data formats, called ‘Simple Features’ (i.e. point, line, polygon, etc.)
- A set of SQL-like instructions for geographic analysis.

The architecture of a spatial database differs from a standard RDBMS not only because it can handle geometry data and manage projections, but also for a larger set of commands that extend standard SQL language (e.g. distance calculations, buffers, overlay, conversion between coordinate systems, etc.).

At the time of writing, spatial databases support the storage of image data, but that support is still relatively limited and under development. As with the hardware and software trends identified in Section 3.2.1, the capabilities of spatial databases will continue to evolve over time. Currently, ESRI's ArcGIS geodatabase can store topological relationships directly in the database, providing support for different kinds of features (objects) and their behaviour (relations with other objects), as well as ways to validate these relations and behaviours. Effectively, this is similar to the functionality offered by traditional DBMSs, but with geospatial data.

Capabilities

Querying a spatial database

A Spatial DBMS provides support for geographic co-ordinate systems and transformations. It also provides storage of the relationships between features, including the creation and storage of topological relationships. As a result one is able to use functions for ‘spatial query’ (exploring spatial relationships). To illustrate, a spatial query using SQL to find all the Thai restaurants within 2 km of a given hotel would look like this:

Spatial query

```
SELECT R.Name  
FROM Restaurants AS R,  
Hotels as H  
WHERE R.Type = "Thai" AND  
H.name = "Hilton" AND  
ST_Intersects(R.Geometry, ST_Buffer(H.Geometry, 2000))
```

In this case the WHERE clause uses the `ST_Intersects` function to perform a spatial join between a 2000 m buffer of the selected hotel and the selected subset of restaurants. The `Geometry` column carries the spatial data.

Summary

Data management and processing functions are central to GIS. This chapter has attempted to provide an overview of DBMS and geodatabase technology. It has examined data management and processing methods and techniques for organizing our spatial and attribute data using GIS and databases.

Traditionally, GIS were more suited for the first and DBMS better for the second purpose. As a result, GIS were often linked to external DBMS for substantial applications or projects requiring more powerful attribute data management capabilities.

Spatial databases are a marriage of GIS and traditional DBMS. They support storage and manipulation of both geometry and attribute data, including spatial queries. The functions and capabilities of spatial databases are constantly improving. In the near future it is likely that we will use spatial databases exclusively for storage of all geometric and attribute data.

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 a laptop. 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 most common technique currently in use? Do you feel better techniques may be available?
3. In Figure 3.2 and Table 3.5 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.)
4. The following is a correct SQL query on the database of Figure 3.2. 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 3.5 (but without showing data values) that demonstrates what the query does.

Chapter 4

Spatial referencing and positioning

In the early days of GIS, users were mainly handling spatially referenced data from a single country. This data was usually derived from paper maps published by the country's mapping organization. Nowadays, GIS users are combining spatial data from a given country with global spatial data sets, reconciling spatial data from published maps with coordinates established with satellite positioning techniques and integrating their spatial data with that from neighbouring countries. To perform these kinds of tasks successfully, GIS users need to understand basic spatial referencing concepts.

This chapter is two parts. In Section 4.1, we discuss the relevance and actual use of reference surfaces, coordinate systems and coordinate transformations. In Section 4.2 we look more closely at satellite-based positioning. The introduction of global positioning techniques has made it possible to unambiguously deter-

mine a position in space. These developments have laid the foundation for the integration of all spatial data within a single global 3D spatial reference system, which we may see emerge within the next 10-15 years.

4.1 Spatial referencing

One of the defining features of GIS is their ability to combine spatially referenced data. A frequently occurring issue is the need to combine spatial data from different sources that use different spatial reference systems. This section provides a broad background of relevant concepts relating to the nature of spatial reference systems and the translation of data from one spatial referencing system into another.

4.1.1 Reference surfaces for mapping

The surface of the Earth is anything but uniform. The oceans can be treated as reasonably uniform, but the surface or topography of the land masses exhibits large vertical variations between mountains and valleys. These variations make it impossible to approximate the shape of the Earth with any reasonably simple mathematical model. Consequently, two main reference surfaces have been established to approximate the shape of the Earth. One reference surface is called the *Geoid*, the other reference surface is the *ellipsoid*. These are illustrated in Figure 4.1. Below, we look at and discuss the respective uses of each of these surfaces.

The Geoid and ellipsoid

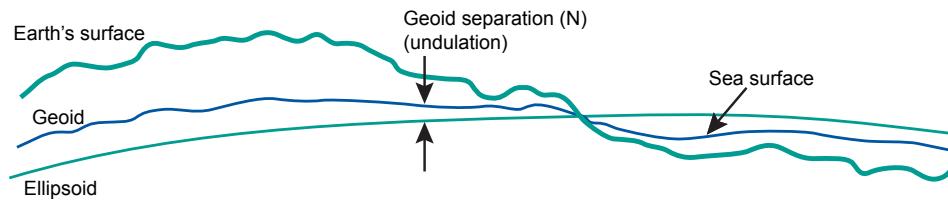


Figure 4.1: The Earth's surface, and two reference surfaces used to approximate it: the Geoid, and a reference ellipsoid. The Geoid separation (N) is the deviation between the Geoid and a reference ellipsoid.

The Geoid and the vertical datum

We can simplify matters by imagining that the entire Earth's surface is covered by water. If we ignore tidal and current effects on this 'global ocean', the resultant water surface is affected only by gravity. This has an effect on the shape of this surface because the direction of gravity—more commonly known as plumb line—is dependent on the mass distribution inside the Earth. Due to irregularities or mass anomalies in this distribution the 'global ocean' results in an undulated surface. This surface is called the Geoid (Figure 4.2). The plumb line through any surface point is always perpendicular to it.

Plumb line

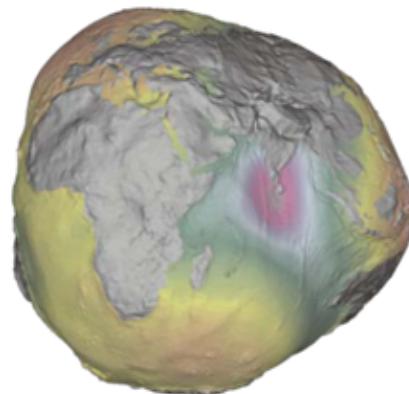


Figure 4.2: The Geoid, exaggerated to illustrate the complexity of its surface.

The Geoid is used to describe *heights*. In order to establish the Geoid as reference for heights, the ocean's water level is registered at coastal places over several years using tide gauges (mareographs). Averaging the registrations largely eliminates variations of the sea level with time. The resulting water level represents an approximation to the Geoid and is called the mean sea level. 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 (Figure 4.3). The result of this process will be the height above local mean sea level for the Enschede point. The height determined with respect to a tide-gauge station is known as the *orthometric height* (height H above the Geoid).

Mean sea level

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. Care must be taken when using heights from another local vertical datum. For example, this might be the case in the border area of adjacent nations. Even within a country, heights may differ depending on to which tide gauge, mean sea level point, they are related. As an example, the mean sea level from the Atlantic to the Pacific coast of the USA increases by 0.6 to 0.7 m. The tide gauge (zero height) of the Netherlands differs -2.34 metres from the tide gauge (zero height) of the neighbouring country Belgium.

Local vertical datums

The local vertical datum is implemented through a levelling network (see Figure 4.3(a)). 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 surveyors 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 (Figure 4.3(b)).

Geodetic levelling

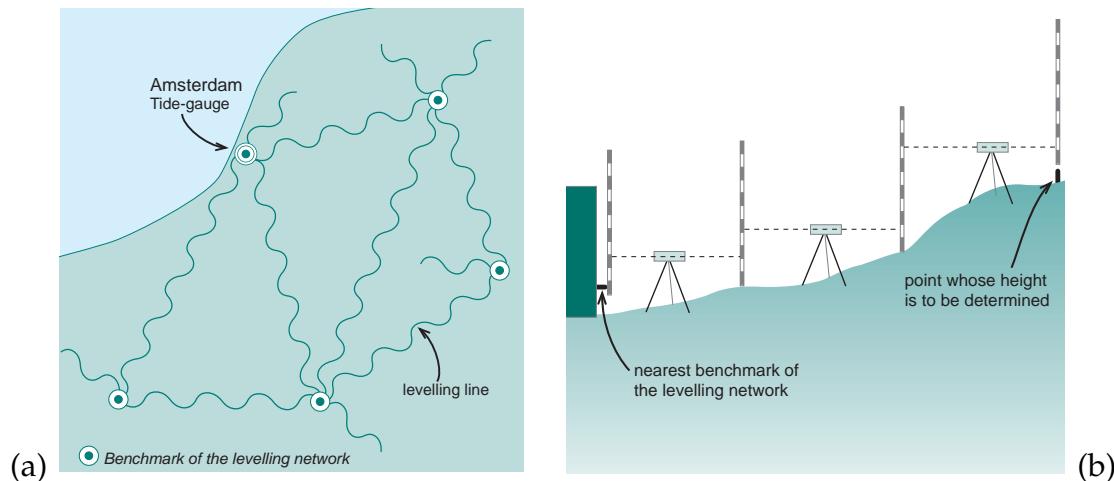


Figure 4.3: A levelling network implements a local vertical datum: (a) network of levelling lines starting from the Amsterdam tide-gauge, showing some of the benchmarks; (b) how the orthometric height (H) is determined for some point, working from the nearest benchmark.

As a result of satellite gravity missions, it is currently possible to determine the height (H) above the Geoid with centimetre level accuracy. It is foreseeable that a global vertical datum may become ubiquitous in the next 10-15 years. If all published maps are also using this global vertical datum by that time, heights will become globally comparable, effectively making local vertical datums redundant for GIS users.

The ellipsoid

Above, we have defined a physical surface, the Geoid, as a reference surface for heights. We also need a reference surface for the description of the *horizontal coordinates* of points of interest. Since we will later project these horizontal coordinates onto a mapping plane, the reference surface for horizontal coordinates requires a mathematical definition and description. The most convenient geometric reference is the *oblate ellipsoid* (Figure 4.4). It provides a relatively simple figure which fits the Geoid to a first order approximation, though for small scale mapping purposes a *sphere* may be used. An ellipsoid is formed when an ellipse is rotated about its minor axis. This ellipse which defines an ellipsoid or *spheroid* is called a meridian ellipse.¹

Oblate ellipsoid

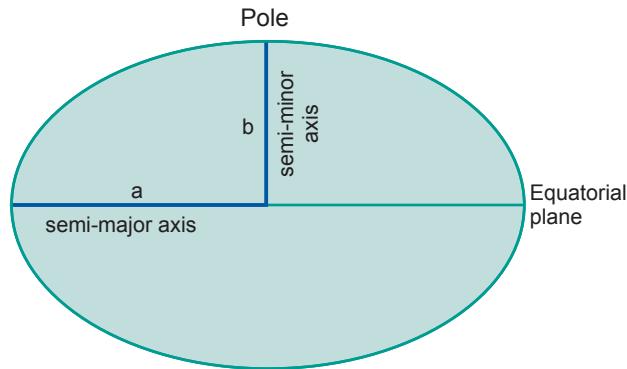


Figure 4.4: An oblate ellipse, defined by its semi-major axis a and semi-minor axis b .

The shape of an ellipsoid may be defined in a number of ways, but in geodetic practice the definition is usually by its semi-major axis and flattening (Fig-

¹Notice that ellipsoid and spheroid are used here to refer to the same thing.

ure 4.4). Flattening f is dependent on both the semi-major axis a and the semi-minor axis b .

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

The ellipsoid may also be defined by its semi-major axis a and its eccentricity e , which is given by:

$$e^2 = \left(1 - \left(\frac{b^2}{a^2}\right)\right) = \frac{(a^2 - b^2)}{a^2} = 2f - f^2.$$

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

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

Many different ellipsoids have been defined. Local ellipsoids have been established to fit the Geoid (mean sea level) well over an area of local interest, which in the past was never larger than a continent. This meant that the differences between the Geoid and the reference ellipsoid could effectively be ignored, allowing accurate maps to be drawn in the vicinity of the datum (Figure 4.5).

Local ellipsoids

With increasing demands for global surveying, work is underway to develop global reference ellipsoids. In contrast to local ellipsoids, which apply only to a

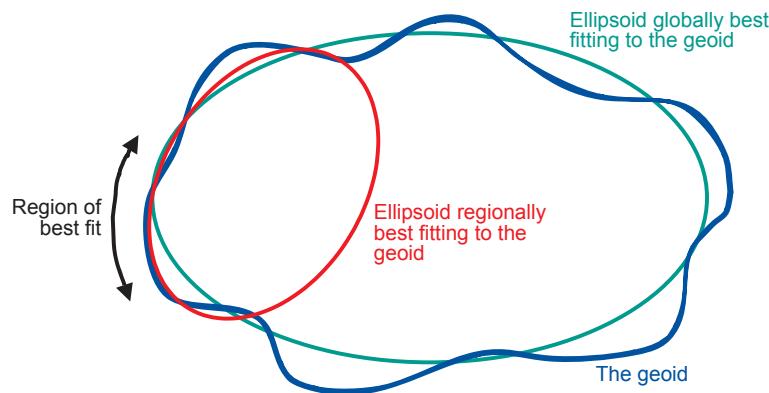


Figure 4.5: The Geoid, a globally best fitting ellipsoid for it, and a regionally best fitting ellipsoid for it, for a chosen region. Adapted from: Ordnance Survey of Great Britain. *A Guide to Coordinate Systems in Great Britain*.

specific country or localised area of the Earth's surface, global ellipsoids approximate the Geoid as a mean earth ellipsoid. The International Union for Geodesy and Geophysics (IUGG) plays a central role in establishing these reference figures.

In 1924, the general assembly of the IUGG in Madrid introduced the ellipsoid determined by Hayford in 1909 as the international ellipsoid. However, according to present knowledge, the values for this ellipsoid give an insufficient approximation. At the general assembly 1967 of the IUGG in Luzern, the 1924 reference system was replaced by the Geodetic Reference System 1967 (GRS 1967). It represents a good approximation (as of 1967) to the mean Earth figure.

For some time, the Geodetic Reference System 1967 was used in the planning of new geodetic surveys. For example, the Australian Datum (1966) and the South American datum (1969) are based upon this ellipsoid. However, at its general assembly 1979 in Canberra the IUGG recognized that the GRS 1967 no longer rep-

Global ellipsoids

resented the size and shape of the Earth to an adequate accuracy. Consequently, it was replaced by the Geodetic Reference System 1980 (GRS80) ellipsoid.

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

Table 4.1: Three global ellipsoids defined by a semi-major axis a , semi-minor axis b , and flattening f . The GRS80 and WGS84 can be considered identical for all practical purposes.

The local horizontal datum

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

Several hundred local horizontal datums exist in the world. The reason is obvious: Different local ellipsoids with varying position and orientation had to be adopted to best fit the local mean sea level in different countries or regions. 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.

A local horizontal datum is realized through a triangulation network. Such a network consists of monumented points forming a network of triangular mesh elements (Figure 4.6). 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.

Triangulation networks

Within this framework, users do not need to start from scratch (i.e. from the fundamental point) in order to determine the geographic coordinates of a new point. They can use the monument of the triangulation network that is closest to the new point. The extension and re-measurement of the network is nowadays done through satellite measurements.



Figure 4.6: The old primary triangulation network in the Netherlands made up of 77 points (mostly church towers). The extension and re-measurement of the network is nowadays done through satellite measurements. Adapted from original figure by 'Dutch Cadastre and Land Registers' now called *het Kadaster*.

The global horizontal datum

Local horizontal datums have been established to fit the Geoid well over the area of local interest, which in the past was never larger than a continent. With increasing demands for global surveying activities are underway to establish global reference surfaces. The motivation is to make geodetic results mutually comparable and to provide coherent results also to other disciplines like astronomy and geophysics.

The most important global (geocentric) spatial reference system for the GIS community is the International Terrestrial Reference System (ITRS). It is a three-dimensional coordinate system with a well-defined origin (the centre of mass of the Earth) and three orthogonal coordinate axes (X , Y , Z). The Z -axis points towards a mean 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 4.7a).

ITRS

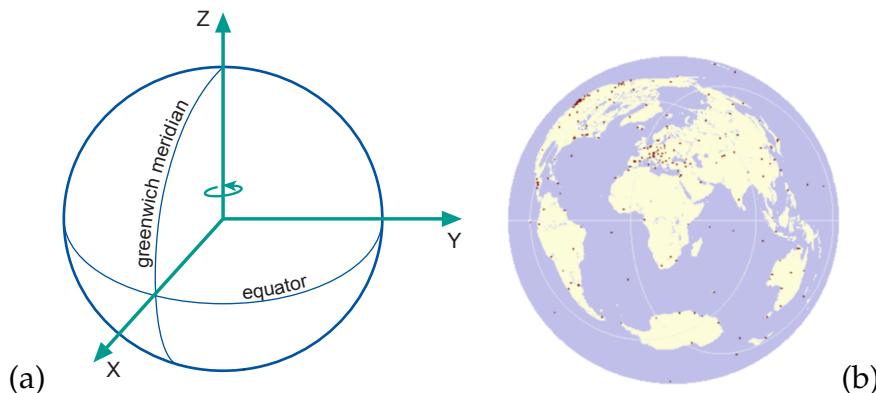


Figure 4.7: (a) The International Terrestrial Reference System (ITRS), and; (b) the International Terrestrial Reference Frame (ITRF) visualized as a distributed set of ground control stations (represented by red points).

The ITRS is realized through the International Terrestrial Reference Frame (ITRF), a distributed set of ground control stations that measure their position continuously using GPS (Figure 4.7b). Constant re-measuring is needed because of the involvement of new control stations and ongoing geophysical processes (mainly tectonic plate motion) that deform the Earth's crust at measurable global, regional and local scales. These deformations cause positional differences in time, and have resulted in more than one realization of the ITRS. Examples are the ITRF96 or the ITRF2000. The ITRF96 was established at the 1st of January, 1997. This means that the measurements use data up to 1996 to fix the geocentric coordinates (X , Y and Z in metres) and velocities (positional change in X , Y and Z in metres per year) at the different stations. The velocities are used to propagate the measurements to other epochs (times). The trend is to use the ITRF everywhere in the world for reasons of global compatibility.

ITRF

GPS uses the World Geodetic System 1984 (WGS84) as its reference system. It has been refined on several occasions and is now aligned with the ITRF to within a few centimetres worldwide. Global horizontal datums, such as the ITRF2000 or WGS84, are also called geocentric datums because they are geocentrically positioned with respect to the centre of mass of the Earth. They became available only recently (roughly after the 1960's), with advances in extra-terrestrial positioning techniques.²

Geocentric datums

Since the size and shape of satellite orbits is directly related to the centre of mass of the Earth, observations of natural or artificial satellites can be used to pinpoint

²Extra-terrestrial positioning techniques include Satellite Laser Ranging (SLR), Lunar Laser Ranging (LLR), Global Positioning System (GPS), and Very Long Baseline Interferometry (VLBI), among others.

the centre of mass of the Earth, and hence the origin of the ITRS.³ This technique can also be used for the realization of the global ellipsoids and datums at the accuracy level required for large-scale mapping.

To implement the ITRF in a region, a densification of control stations is needed to ensure that there are enough coordinated reference points available in the region. These control stations are equipped with 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.

We can easily transform ITRF coordinates (X , Y and Z in metres) into geographic coordinates (ϕ , λ , h) with respect to the GRS80 ellipsoid without the loss of accuracy. However, the ellipsoidal height h , obtained through this straightforward transformation, has no physical meaning and does not correspond to intuitive human perception of height. We therefore use the height H , above the Geoid (see Figure 4.8). It is foreseeable that global 3D spatial referencing, in terms of (ϕ , λ , H), could become ubiquitous in the next 10–15 years. If all published maps are also globally referenced by that time, the underlying spatial referencing concepts will become transparent and hence redundant for GIS users.

3D spatial referencing

Hundreds of existing local horizontal and vertical datums are still relevant because they form the basis of map products all over the world. For the next few years we will be required to deal with both local and global datums until the former are eventually phased out. During the transition period, we will require tools to transform coordinates from local horizontal datums to a global hori-

³In the case of an idealized spherical Earth it is one of the focal points of the elliptical orbits.

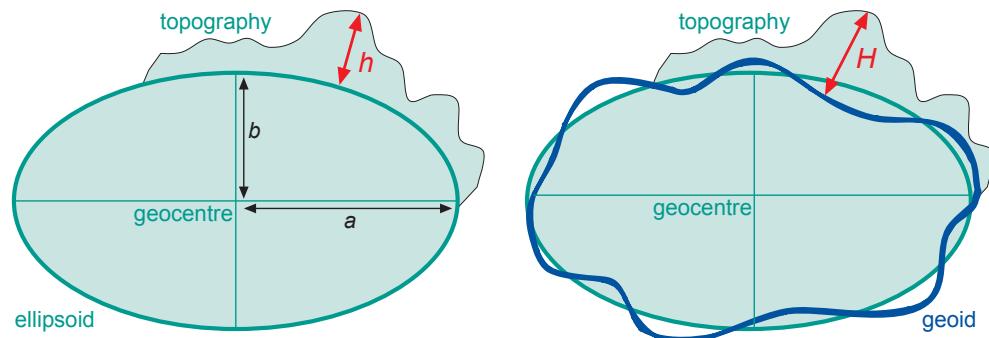


Figure 4.8: Height h above the geocentric ellipsoid, and height H above the Geoid. The first is measured orthogonal to the ellipsoid, the second orthogonal to the Geoid.

horizontal datum and vice versa (see Section 4.1.4). The organizations that usually develop transformation tools and make them available to the user community are provincial or National Mapping Organizations (NMOs) and cadastral authorities.

4.1.2 Coordinate systems

As mentioned before, the special nature of spatial data obviously lies in it being spatially referenced. Different kinds of coordinate systems are used to position data in space. Here we distinguish between *spatial* and *planar* coordinate systems. Spatial (or global) coordinate systems are used to locate data either on the Earth's surface in a 3D space, or on the Earth's reference surface (ellipsoid or sphere) in a 2D space. Below we discuss the geographic coordinate system in a 2D and 3D space and the geocentric coordinate system, also known as the 3D Cartesian coordinate system. Planar coordinate systems on the other hand are used to locate data on the flat surface of the map in a 2D space. We will discuss the 2D Cartesian coordinate system and the 2D polar coordinate system.

Spatial and planar
coordinate systems

2D Geographic coordinates (ϕ, λ)

The most widely used global coordinate system consists of lines of geographic *latitude* (phi or ϕ or φ) and *longitude* (lambda or λ). Lines of equal latitude are called parallels. They form circles on the surface of the ellipsoid⁴. Lines of equal longitude are called meridians and they form ellipses (meridian ellipses) on the ellipsoid. (Figure 4.9)

Latitude and longitude

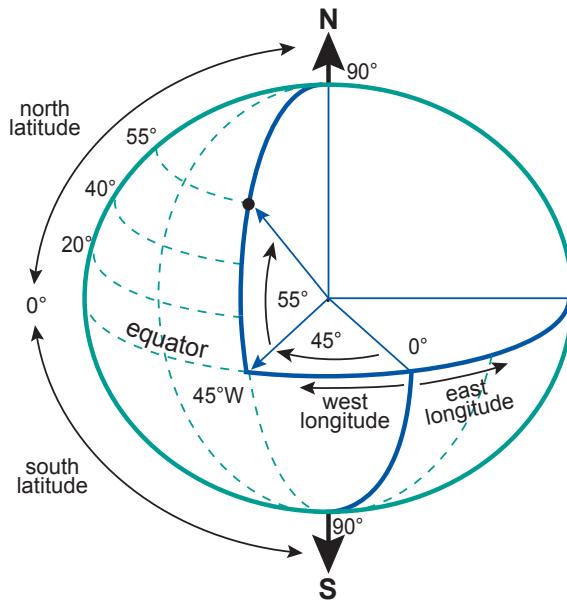


Figure 4.9: The latitude (ϕ) and longitude (λ) angles represent the 2D geographic coordinate system.

The latitude (ϕ) of a point P (Figure 4.10) is the angle between the ellipsoidal

⁴The concept of geographic coordinates can also be applied to a sphere.

normal through P' and the equatorial plane. Latitude is zero on the equator ($\phi = 0^\circ$), and increases towards the two poles to maximum values of $\phi = +90^\circ$ ($N\ 90^\circ$) at the North Pole and $\phi = -90^\circ$ ($S\ 90^\circ$) at the South Pole.

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

Latitude and longitude represent the geographic coordinates (ϕ, λ) of a point P' (Figure 4.10) with respect to the selected reference surface. They are always given in angular units. For example, the coordinates for City hall in Enschede are:⁵

$$\phi = 52^\circ 13' 26.2'' N, \lambda = 6^\circ 53' 32.1'' E$$

The graticule on a map represents the projected position of the geographic coordinates (ϕ, λ) at constant intervals, or in other words the projected position of selected meridians and parallels (Figure 4.13). The shape of the graticule depends largely on the characteristics of the map projection and the scale of the map.

Graticule

⁵This latitude and longitude refers to the Amersfoort datum. The use of a different reference surface will result in a different latitude and longitude angle.

3D Geographic coordinates (ϕ, λ, h)

3D geographic coordinates (ϕ, λ, h) are obtained by introducing the ellipsoidal height h to the system. The ellipsoidal height (h) of a point is the vertical distance of the point in question above the ellipsoid. It is measured in distance units along the ellipsoidal normal from the point to the ellipsoid surface. 3D geographic coordinates can be used to define a position on the surface of the Earth (point P in Figure 4.10).

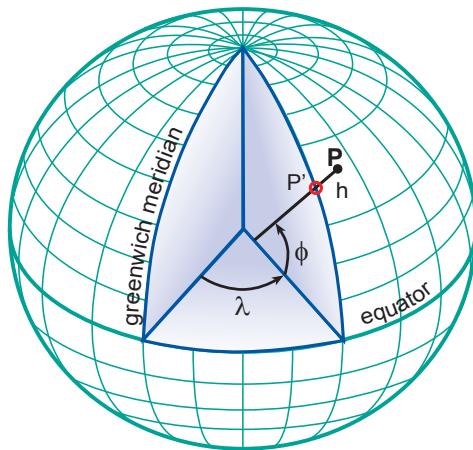


Figure 4.10: The latitude (ϕ) and longitude (λ) angles and the ellipsoidal height (h) represent the 3D geographic coordinate system.

3D Geocentric coordinates (X, Y, Z)

An alternative method of defining a 3D position on the surface of the Earth is by means of geocentric coordinates (X, Y, Z), also known as *3D Cartesian coordinates*. The system has its origin at the mass-centre of the Earth with the X and Y axes in the plane of the equator. The X -axis passes through the meridian of Greenwich, and the Z -axis coincides with the Earth's axis of rotation. The three axes are mutually orthogonal and form a right-handed system. Geocentric coordinates can be used to define a position on the surface of the Earth (point P in Figure 4.11).

It should be noted that the rotational axis of the earth changes its position over time (referred to as *polar motion*). To compensate for this, the mean position of the pole in the year 1903 (based on observations between 1900 and 1905) has been used to define the so-called 'Conventional International Origin' (CIO).

Polar motion

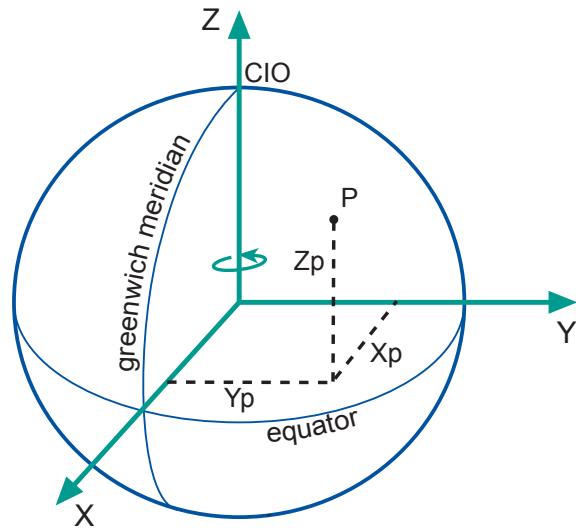


Figure 4.11: An illustration of the 3D geocentric coordinate system (see text for further explanation).

2D Cartesian coordinates (X, Y)

A flat map has only two dimensions: width (left to right) and length (bottom to top). Transforming the three dimensional Earth into a two-dimensional map is subject of map projections and coordinate transformations (Section 4.1.3 and Section 4.1.4). Here, like in several other cartographic applications, *two-dimensional Cartesian coordinates* (x, y), also known as *planar rectangular coordinates*, are used to describe the location of any point unambiguously.

The 2D Cartesian coordinate system is a system of intersecting perpendicular lines, which contains two principal axes, called the X - and Y -axis. The horizontal axis is usually referred to as the X -axis and the vertical the Y -axis (Note that the X -axis is also sometimes called Easting and the Y -axis the Northing). The intersection of the X and Y -axis forms the *origin*. The plane is marked at intervals by equally spaced coordinate lines, called the *map grid*.

Eastings, Northings and
map grid

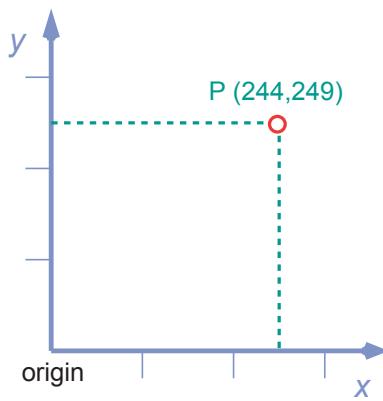


Figure 4.12: An illustration of the 2D Cartesian coordinate system (see text for further explanation).

Given two numerical coordinates x and y for point P , one can now precisely and objectively specify any location P on the map (Figure 4.12).

Normally, the coordinates $x=0$ and $y=0$ are given to the origin. However, sometimes large positive values are added to the origin coordinates. This is to avoid negative values for the x and y coordinates in case the origin of the coordinate system is located inside the area of interest. The point which then has the coordinates $x=0$ and $y=0$ is called the *false origin*.

False origin

An example is the coordinate system used in the Netherlands. It is called Rijksdriehoekstelsel (RD). The system is based on the azimuthal stereographic projection (see Section 4.1.3) and the Bessel ellipsoid is used as reference surface. The origin of the coordinate system has been shifted (false origin) from the projection centre (Amersfoort) towards the south-west to avoid negative coordinates inside the country (see Figure 4.13).

The grid on a map represents lines having constant 2D Cartesian coordinates (Figure 4.13). It is almost always a rectangular system and is used on large and medium scale maps to enable detailed calculations and positioning. The map grid is usually not used on small scale maps (about one to a million or smaller). Scale distortions that result from transforming the Earth's curved surface to the map plane are so great on small-scale maps that detailed calculations and positioning are difficult.

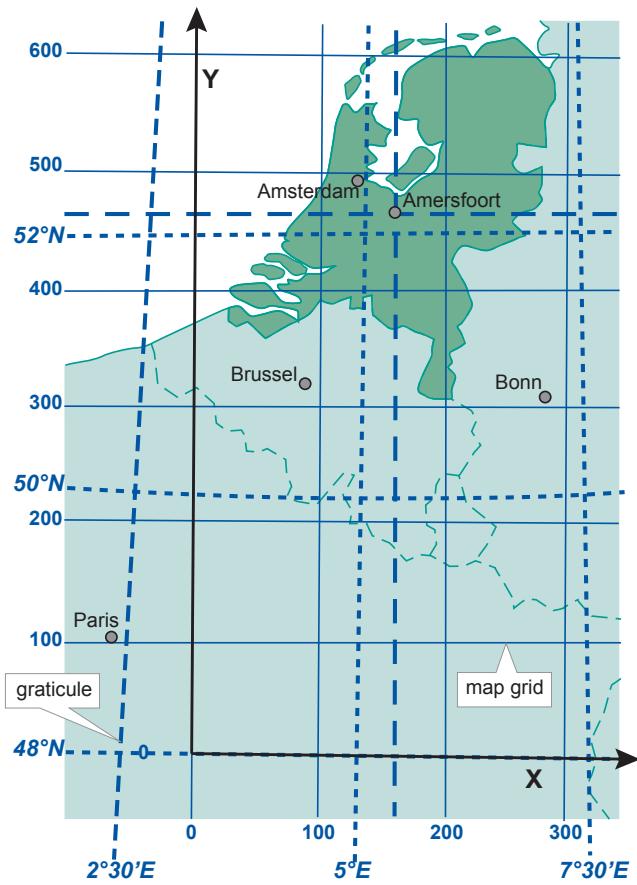


Figure 4.13: The co-ordinate system of the Netherlands represented by the map grid and the graticule. The origin of the coordinate system has been shifted (false origin) from the projection centre (Amersfoort) towards the South-West.

2D Polar coordinates (α, d)

Another possibility of defining a point in a plane is by polar coordinates. This is the distance d from the origin to the point concerned and the angle α between a fixed (or zero) direction and the direction to the point. The angle α is called *azimuth* or *bearing* and is measured in a clockwise direction. It is given in angular units while the distance d is expressed in length units.

Bearing or azimuth

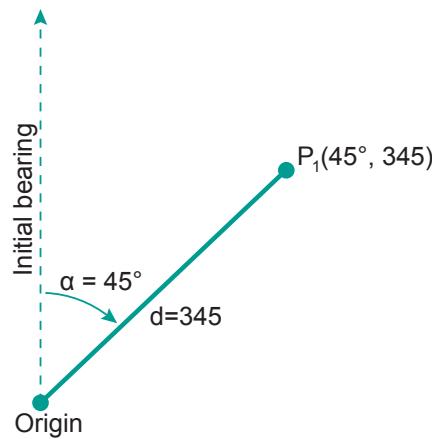


Figure 4.14: An illustration of the 2D Polar coordinate system (see text for further explanation).

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

Polar coordinates are often used in land surveying. For some types of surveying instruments it is advantageous to make use of this coordinate system. The

development of precise remote distance measurement techniques has led to the virtually universal preference for the polar coordinate method in detailed surveys.

4.1.3 Map projections

Maps are one of the world's oldest types of document. For quite some time it was thought that our planet was *flat*, and during those days, a map simply was a miniature representation of a part of the world. Now that we know that the Earth's surface is curved in a specific way, we know that a map is in fact a flattened representation of some part of the planet. The field of map projections concerns itself with the ways of translating the curved surface of the Earth into a flat map.

A *map projection* is a mathematically described technique of how to represent the Earth's curved surface on a flat map.

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 for large-scale mapping is usually an oblate ellipsoid, and for small-scale mapping, a sphere.⁶ Mapping onto a 2D mapping plane means transforming each point on the reference surface with geographic coordinates (ϕ, λ) to a set of Cartesian coordinates (x, y) representing positions on the map plane (Figure 4.15).

The actual mapping cannot usually be visualized as a true geometric projection, directly onto the mapping plane (Figure 4.15). This is achieved through mapping equations. A *forward mapping equation* transforms the geographic coordi-

Mapping equations

⁶In practice, maps at scale 1:1,000,000 or smaller can use the mathematically simpler sphere without the risk of large distortions. At larger scales, the more complicated mathematics of ellipsoids are needed to prevent these distortions in the map.

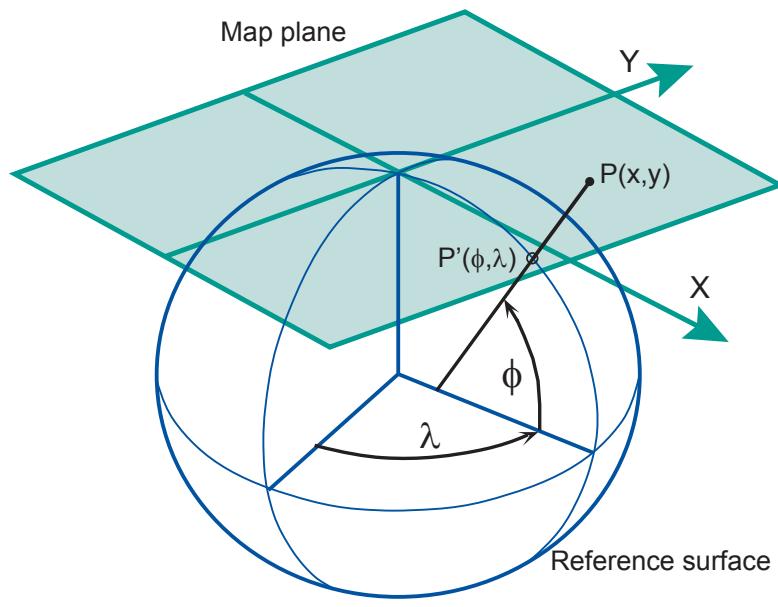


Figure 4.15: Example of a map projection where the reference surface with geographic coordinates (ϕ, λ) is projected onto the 2D mapping plane with 2D Cartesian coordinates (x, y) .

nates (ϕ, λ) of a point on the curved reference surface to a set of planar Cartesian coordinates (x, y) , representing the position of the same point on the map plane:

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

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

coordinates (ϕ, λ) on the curved reference surface:

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

An example is the mapping equations used for the Mercator projection (spherical assumption) [51]. The forward mapping equation for the Mercator projection is:⁷

$$x = R(\lambda - \lambda_0)$$

$$y = R(\ln(\tan(\frac{\pi}{4} + \frac{\phi}{2})))$$

The inverse mapping equation for the Mercator projection is:

$$\phi = \frac{\pi}{2} - 2 \arctan(e^{\frac{-y}{R}})$$

$$\lambda = \frac{x}{R} + \lambda_0$$

⁷The equations are considerably more complicated than those introduced here when an ellipsoid is used as reference surface. R is the radius of the spherical reference surface at the scale of the map; ϕ and λ are given in radians; λ_0 is the central meridian of the projection; $e = 2.7182818$, the base of the natural logarithms, not the eccentricity.

Classification of map projections

Hundreds of map projections have been developed, each with its own specific qualities. These qualities in turn make resulting maps useful for certain purposes. By definition, any map projection is associated with scale 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. The amount and which kind of distortions a map will have depends on the type of the map projection that has been selected.

Some map projections can be visualized as true geometric projections directly onto the mapping plane, in which case we call it an azimuthal projection, 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 conical, and cylindrical, respectively. Figure 4.16 shows the surfaces involved in these three classes of projections.

The planar, conical, and cylindrical surfaces in Figure 4.16 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 4.17. Then, the reference surface is intersected along one closed line (plane) or two closed lines (cone and cylinder). Secant map surfaces are used to reduce or average out scale errors because the line(s) of intersection are not distorted on the map.

In the geometric depiction of map projections in Figures 4.16 and 4.17, the symmetry axes of the plane, cone and cylinder coincide with the rotation axis of the

Scale distortions

Intermediate surfaces

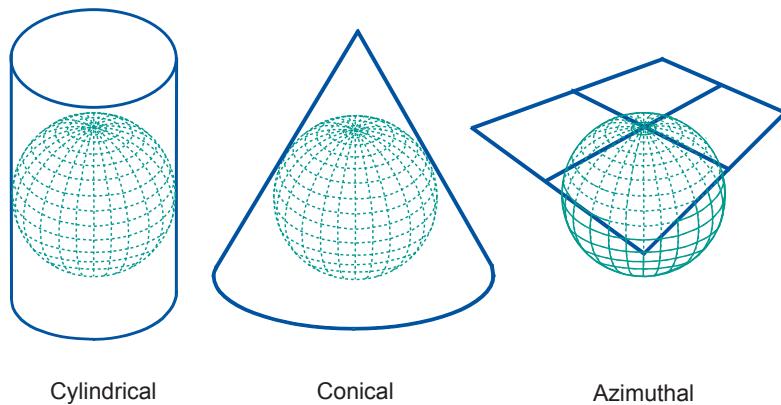


Figure 4.16: Classes of map projections

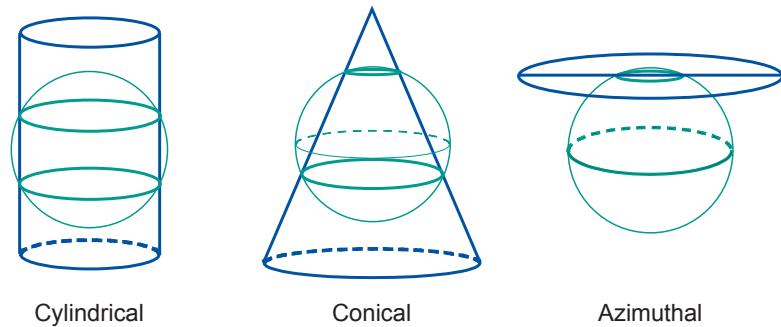


Figure 4.17: Three secant projection classes

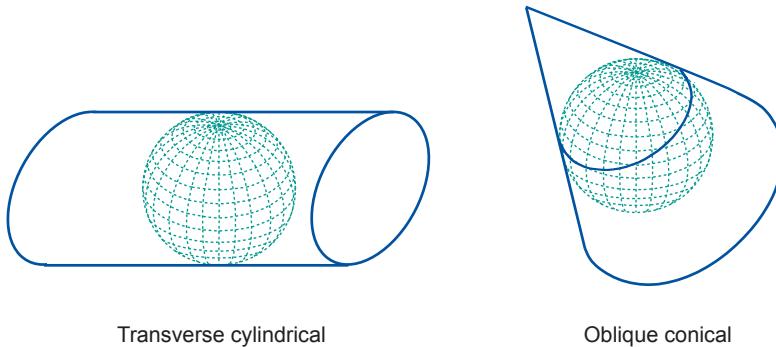


Figure 4.18: A transverse and an oblique projection

ellipsoid or sphere, i.e. a line through N and S pole. In this case, the projection is said to be a *normal projection*. The other cases are *transverse projections* (symmetry axis in the equator) and *oblique projections* (symmetry axis is somewhere between the rotation axis and equator of the ellipsoid or sphere). These cases are illustrated in Figure 4.18.

Normal, transverse, and oblique projections

UTM

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

Normal cylindrical projections are typically used to map the world in its entirety. Conical projections are often used to map the different continents, while the normal azimuthal projection may be used to map the polar areas. Transverse and oblique aspects of many projections can be used for most parts of the world.

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

So far, we have not specified *how* the curved horizontal reference surface is projected onto the plane, cone or cylinder. *How* this is done 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 in the map are identical to the angles between the original lines on the curved reference surface. This means that angles (with short sides) and shapes (of small areas) are shown correctly on the map.
- In an *equal-area* (equivalent) map projection the areas in the map are - identical to the areas on the curved reference surface (taking into account the map scale), which means that areas are represented correctly on the map.
- In an *equidistant* map projection the length of particular lines in the map are the same as the length of the original lines on the curved reference surface (taking into account the map scale).

Distortion properties

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

The most appropriate type of distortion property for a map depends largely on the purpose for which it will be used. Conformal map projections represent angles correctly, but as the region becomes larger, they show considerable area distortions (Figure 4.19). Maps used for the measurement of angles (e.g. aeronautical charts, topographic maps) often make use of a conformal map projection such as the UTM projection.

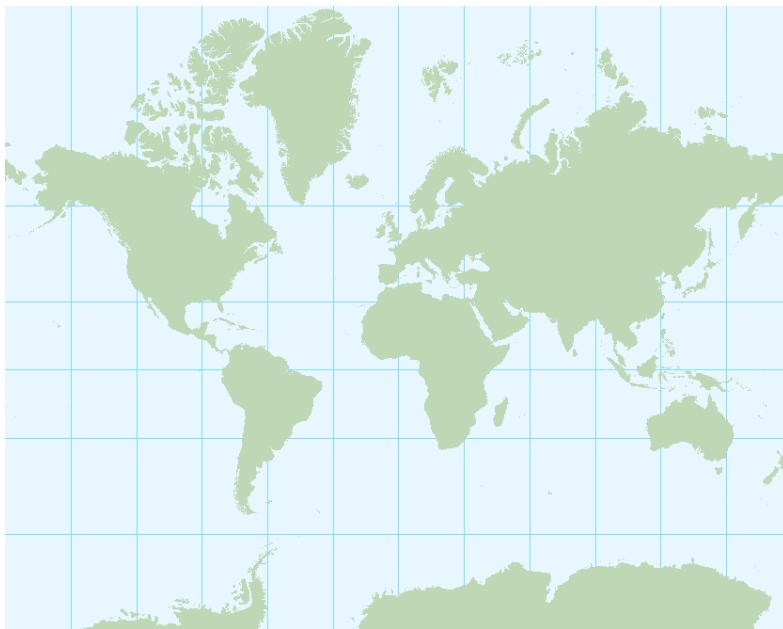


Figure 4.19: The Mercator projection, a cylindrical map projection with a conformal property. The area distortions are significant towards the polar regions.

Equal-area projections on the other hand, represent areas correctly, but as the region becomes larger, it shows considerable distortions of angles and consequently shapes (Figure 4.20). Maps which are to be used for measuring area

(e.g. distribution maps) often make use of an equal-area map projection.

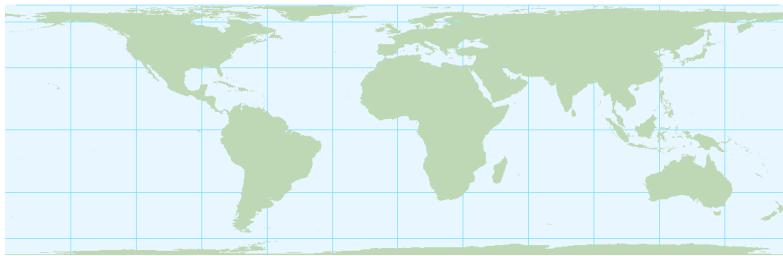


Figure 4.20: The cylindrical equal-area projection, a cylindrical map projection with an equal-area property. The shape distortions are significant towards the polar regions.

The equidistant property is achievable only to a limited degree. That is, true distances can be shown only from one or two points to any other point on the map or in certain directions. If a map is true to scale along the meridians (i.e. no distortion in North-South direction) we say that the map is *equidistant along the meridians* (e.g. the equidistant cylindrical projection) (Figure 4.21). If a map is true to scale along all parallels we say the map is *equidistant along the parallels* (i.e. no distortion in East-West direction). Maps which require reasonable area and angle distortions (several thematic maps) often make use of an equidistant map projection.

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. (This would amount to the projection of Figure 4.17, middle.)

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

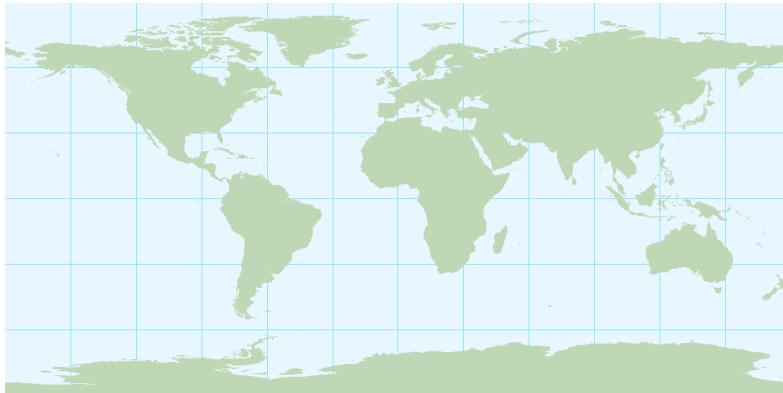


Figure 4.21: The equidistant cylindrical projection (also called Plate Carrée projection), a cylindrical map projection with an equidistant property. The map is equidistant (true to scale) along the meridians. Both shape and area are reasonably well preserved.

first publisher). For example, the ‘conformal conic projection with two standard parallels’ is also referred to as ‘Lambert’s conical projection’ [24].

4.1.4 Coordinate transformations

Map and GIS users are mostly confronted in their work with transformations from one two-dimensional coordinate system to another. This includes the transformation of polar coordinates delivered by the surveyor into Cartesian map coordinates or the transformation from one 2D Cartesian (x, y) system of a specific map projection into another 2D Cartesian (x', y') system of a defined map projection.

Datum transformations are transformations from a 3D coordinate system (i.e. horizontal datum) into another 3D coordinate system. These kinds of transformations are also important for map and GIS users. They are usually collecting spatial data in the field using satellite navigation technology and need to represent this data on published map on a local horizontal datum.

We may relate an unknown coordinate system to a known coordinate system on the basis of a set of selected points whose coordinates are known in both systems. These points may be ground control points (GCPs) or common points such as corners of houses or road intersections, as long as they have known coordinates in both systems. Image and scanned data are usually transformed by this method. The transformations may be conformal, affine, polynomial, or of another type, depending on the geometric errors in the data set. These type of 2D Cartesian transformations are not covered in this textbook, but are discussed in *Principles of Remote Sensing* [53].

2D Polar to 2D Cartesian transformations

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

$$x = d(\sin(a))$$

$$y = d(\cos(a))$$

The inverse equation is:

$$a = \tan^{-1}\left(\frac{x}{y}\right)$$

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

A more realistic case makes use of a translation and a rotation to transform one system to the other.

Changing 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) into target projection coordinates (x', y') . The first equation takes us from a projection A into geographic coordinates. The second takes us from geographic coordinates (ϕ, λ) to another map projection B . These principles are illustrated in Figure 4.22.

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 the location of objects in a 3D GIS.

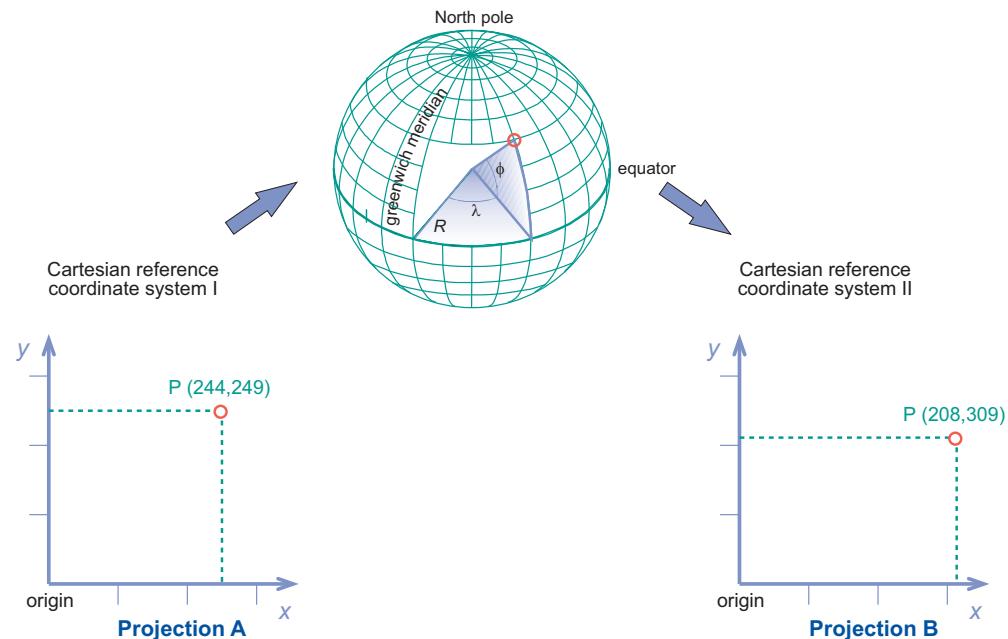


Figure 4.22: The principle of changing from one map projection into another.

Datum transformations

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

Suppose we wish to transform spatial data from the UTM projection to the Dutch RD system, and that the data in the UTM system are related to the European Datum 1950 (ED50), while the Dutch RD system is based on the Amersfoort datum. In this example the change of map projection should be combined with a datum transformation step for a perfect match. This is illustrated in Figure 4.23.

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

Mathematically a datum transformation is feasible via the geocentric coordinates (x, y, z) , or directly by relating the geographic coordinates of both datum systems. The latter relates the ellipsoidal latitude (ϕ) and longitude (λ), and

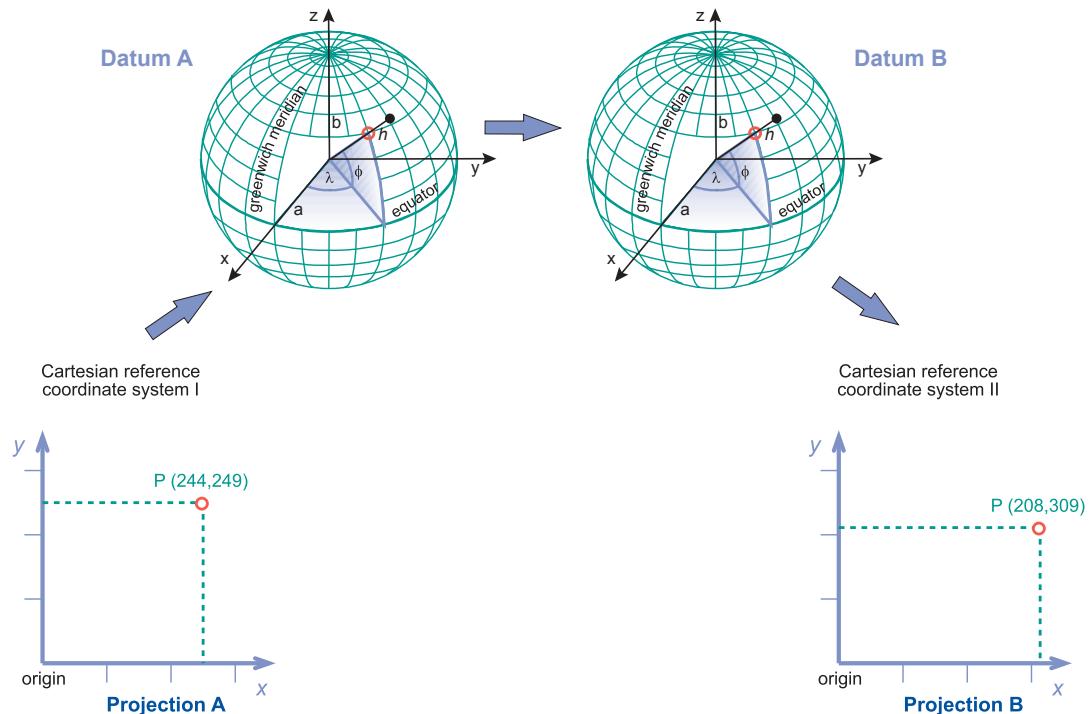


Figure 4.23: The principle of changing from one projection into another, combined with a datum transformation from datum A to datum B.

possibly also the ellipsoidal height (h), of both datum systems [28].

We can easily transform geographic coordinates (ϕ, λ, h) into geocentric coordinates (x, y, z) , and the other way around. The datum transformation via the geocentric coordinates implies a 3D similarity transformation. Essentially, this is a transformation between two orthogonal 3D Cartesian spatial reference frames together with some elementary tools from adjustment theory. The transformation is usually expressed with seven parameters: three rotation angles (α, β, γ) , three origin shifts (X_0, Y_0, Z_0) and 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 seven transformation parameters and a measure of the likely error of the estimate.

Datum transformation parameters have to be estimated on the basis of a set of selected points whose coordinates are known in both datum systems. If the coordinates of these 5 points are not correct—often the case for points measured on a local datum system—the estimated parameters may be inaccurate. As a result the datum transformation will be inaccurate. This is often the case when we transform coordinates from a local horizontal datum to a global geocentric datum. The coordinates in the local horizontal datum may be distorted by several tens of metres because of the inherent inaccuracies of the measurements used in the triangulation network. These inherent inaccuracies are also responsible for another complication: the transformation parameters are not unique. Their estimate will depend on which particular common points are chosen, and they also will depend on whether all seven transformation parameters, or only a sub-set of them, are estimated.

Datum transformation
parameters

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

<i>Parameter</i>		<i>National set</i>	<i>Provincial set</i>	<i>NIMA set</i>
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 4.2: Three different sets of datum transformation parameters from three different organizations for transforming a point from ITRF to the Potsdam datum.

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:

1. The set provided by the federal mapping organization of Germany (labelled 'National set' in Table 4.2) was calculated using common points distributed throughout Germany. This set contains all seven parameters and is valid for all of Germany.
2. The set provided by the mapping organization of Baden-Württemberg (labelled 'Provincial set' in Table 4.2) 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.
3. The set provided by the National Imagery and Mapping Agency (NIMA) of the USA (labelled 'NIMA set' in Table 4.2) has been calculated using

common points distributed throughout Germany and based on the ITRF. This set contains a coordinate shift only (no rotations, and scale equals unity). It is valid for all of Germany.

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 m, 671,428.74 m, 4,774,958.21 m).

The three sets of transformed coordinates in the Potsdam datum are given in Table 4.3. 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.

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

Table 4.3: Three sets of transformed coordinates for a point in the state of Baden-Württemberg.

4.2 Satellite-based positioning

The previous section has noted the importance of satellites in spatial referencing. Satellites have allowed us to realize geocentric reference systems, and increase the level of spatial accuracy substantially. They are critical tools in geodetic engineering for the maintenance of the ITRF. They also play a key role in mapping, surveying, and in a growing number of applications requiring positioning techniques. Nowadays, for fieldwork that includes spatial data acquisition, the use of satellite-based positioning is considered indispensable.

Satellite-based positioning was developed and implemented to address military needs, somewhat analogously to the early development of the internet. The technology is now widely available for civilians use. The requirements for the development of the positioning system were:

- Suitability for all kinds of military use: ground troops and vehicles, aircraft and missiles, ships;
- Requiring only low-cost equipment with low energy consumption at the receiver end;
- Provision of results in real time for an unlimited number of users concurrently;
- Support for different levels of accuracy (military versus civilian);
- Around-the-clock and weather-proof availability;
- Use of a single geodetic datum;

- Protection against intentional and unintentional disturbance, for instance, through a design allowing for redundancy.

A satellite-based positioning system set-up involves implementation of three hardware segments:

1. The *space segment*, i.e. the satellites that orbit the Earth, and the radio signals that they emit,
2. The *control segment*, i.e. the ground stations that monitor and maintain the space segment components, and
3. The *user segment*, i.e. the users with their hard- and software to conduct positioning.

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

In the remainder of this section, we discuss some of the fundamentals of satellite-based positioning, having in mind especially the geoscientist that wants to make use of it.

4.2.1 Absolute positioning

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

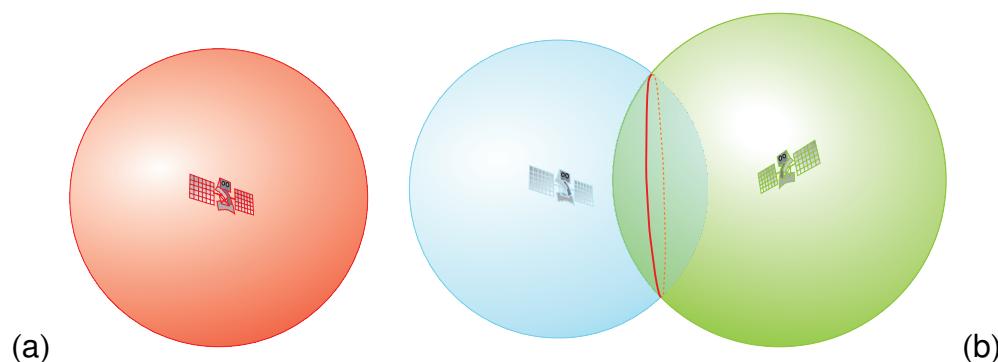
1. A satellite, equipped with a clock, at a specific moment sends a radio message that includes
 - (a) the *satellite identifier*,
 - (b) its *position in orbit*, and
 - (c) its *clock reading*.
2. A receiver on or above the planet, also equipped with a clock, receives the message slightly later, and reads its own clock.
3. From the time delay observed between the two clock readings, and knowing the speed of radio transmission through the medium between (satellite) sender and receiver, the receiver can compute the distance to the sender, also known as the satellite's *pseudorange*.

The *pseudorange* of a satellite with respect to a receiver, is its apparent distance to the receiver, computed from the time delay with which its radio signal is received.

Such a computation determines the position of the receiver to be on a sphere of radius equal to the computed pseudorange (refer to Figure 4.24(a)). If the

receiver instantaneously would do the same with a message of another satellite that is positioned elsewhere, the position of the receiver is restricted to another sphere. The intersection of the two spheres, which have different centres, determines a circle as the set of possible positions of the receiver (refer to Figure 4.24(b)). If a third satellite message is taken into consideration, the intersection of three spheres determines at most two positions, one of which is the actual position of the receiver. In most, if not all, practical situations where two positions result, one of them is a highly unlikely position for a signal receiver. The overall procedure is known as *trilateration*: the determination of a position based on three distances.

Trilateration

**Figure 4.24:**

Pseudorange positioning:
(a) With just one satellite the position is determined by a sphere, (b) With two satellites, it is determined by the intersection of two spheres, a circle. Not shown: with three satellites, it is the intersection of three spheres.

It would appear therefore that the signals of three satellites would suffice to determine a *positional fix* for our receiver. In theory this is true, but in practice it is not. The reason is that we have made the assumption that all satellite clocks as well as our receiver clock are fully synchronized, where in fact they are not. The satellite clocks are costly, high-precision, atomic clocks that we can consider synchronized for the time being, but the receiver typically has a far cheaper, quartz

Clock bias

clock that is not synchronized with the satellite clocks. This brings into play an additional unknown parameter, namely the synchronization bias of the receiver clock, i.e. the difference in time reading between it and the satellite clocks.

Our set of unknown variables has now become $(X, Y, Z, \Delta t)$ representing a 3D position and a clock bias. By including the information obtained from a fourth satellite message, we can solve the problem (see Figure 4.25). This will result in the determination of the receiver's actual position (X, Y, Z) , as well as its receiver clock bias Δt , and if we correct the receiver clock for this bias we effectively turn it into a high-precision, atomic clock as well!

3D positioning

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

Another application is in the positioning of mobile phone users making an emergency call. Often the caller does not know their location accurately. The telephone company can trace back the call to the receiving transmitter mast, but this may be servicing an area with a radius of 300 m to 6 km. That is too inaccurate a position for an emergency ambulance to go to. However, if all masts in the telephony network are equipped with a satellite positioning receiver (and thus, with a very good, synchronized clock) the time of reception of the call at each mast can be recorded. The *time difference of arrival* of the call between two nearby masts determines a hyperbola on the ground of possible positions of the caller; if the call is received on three masts, we would have two hyperbolas, allowing intersection, and thus 'hyperbolic positioning'. With current technology

the (horizontal) accuracy would be better than 30 m.

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

2D positioning mode

If a 3D fix has already been obtained, the receiver simply assumes that the height above the ellipsoid has not changed since the last 3D fix. If no fix had yet been obtained, the receiver assumes that it is positioned at the geocentric ellipsoid adopted by the positioning system, i.e. at height $h=0$.⁸ In the receiver computations, the ellipsoid fills the slot of the missing fourth satellite sphere, and the unknown variables can therefore still be determined. Clearly in both of these cases, the assumption for this computation is flawed and the positioning results in 2D mode will be unreliable—much more so if no previous fix had been obtained and one’s receiver is not at all near the surface of the geocentric ellipsoid.

⁸Any receiver is capable of transforming a triad (X, Y, Z) , using a straightforward mathematical transformation, into an equivalent triad (ϕ, λ, h) , where h is the height above the geocentric ellipsoid.

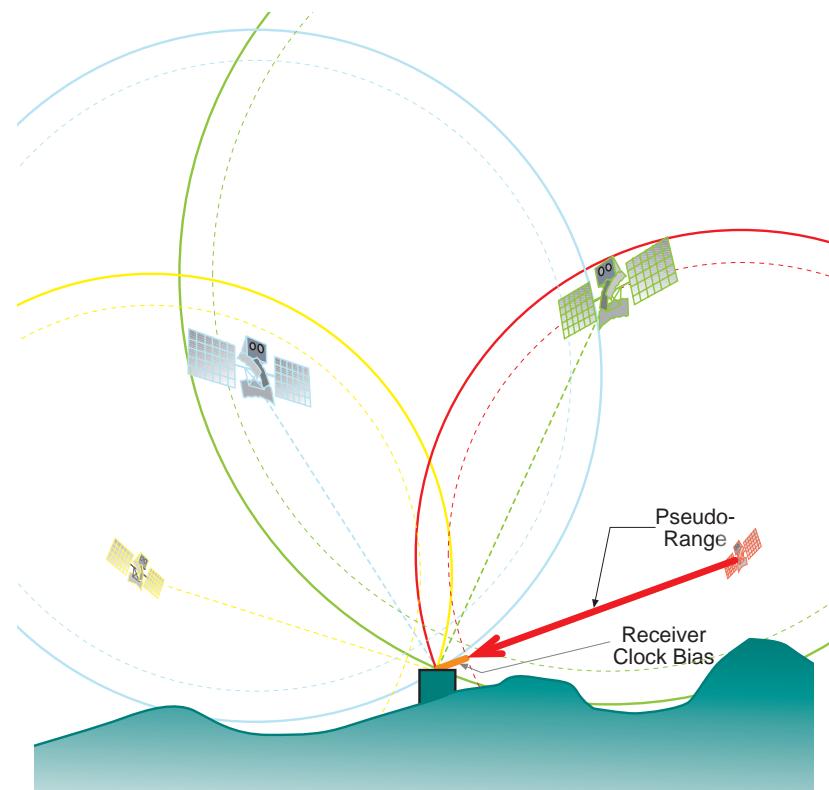


Figure 4.25: Four satellites are needed to obtain a 3D position fix. Pseudoranges are indicated for each satellite as dashed circles representing a sphere, as well as the actual range as a normal circle, being the pseudorange plus the range error caused by receiver clock bias.

Time, clocks and world time

During most of human history, the determination of time and position have gone hand in hand. This was probably true of many civilizations in Asia and Arabia before the Christian calendar as witnessed by remnants of various time keeping constructions, as well as for early civilizations in Latin America; it was certainly true for the European seafarer explorers of the 15th through to the 18th century. While latitude was determined with a sextant from the position of the Sun in the sky, they carried clocks with them to determine the longitude of their position. Early ship clocks were notoriously unreliable, having a drift of multiple seconds a day, which could result in positional error of a few kilometres.

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

Greenwich mean time

GMT was later replaced by Universal Time (UT), a system still based on meridian crossings of stars, but now of far away quasars as this provides more accuracy than that of the Sun. It is still the case that the rotational velocity of our planet is not constant and the length of a solar day is increasing. So UT is not a perfect system either. It continues to be used for civil clock time, but it is officially now replaced by International Atomic Time (TAI). UT actually has various

versions, amongst which are UT0, UT1 and UTC. UT0 is the Earth rotational time observed in some location. Because the Earth experiences polar motion as well, UT0 differs between locations. If we correct for polar motion, we obtain UT1, which is identical everywhere. It is still a somewhat erratic clock because of the earlier mentioned varying rotational velocity of the planet. The uncertainty is about 3 msec per day.

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

In recent years we have learned to measure distance, therefore also position, with clocks using satellite signals. The conversion factor is the speed of light, approximately $3 \cdot 10^8$ m/s in vacuum. No longer can multiple seconds of clock bias be allowed, and this is where atomic clocks come in. They are very accurate time keepers, based on the exactly known frequency with which specific atoms (Cesium, Rubidium and Hydrogen) make discrete energy state jumps. Positioning satellites usually have multiple clocks on board; ground control stations have even better quality atomic clocks.

Atomic clocks, however, are not flawless: their timing tends to drift away from true time somewhat, and they too need to be corrected. The drift, and the change

UTC

Atomic clocks

in drift over time, are monitored, and are part of the satellite's navigation message, so that they can be corrected for.

4.2.2 Errors in absolute positioning

Before we continue discussing other modes of satellite-based positioning, let us take a close look at the potential for error in absolute positioning. Receiver users are required to be sufficiently familiar with the technology in order to avoid true operating blunders such as bad receiver placement or incorrect receiver software settings, which can render the results virtually useless. We will skip over many of the physical and mathematical details underlying these errors, but they are mentioned here to raise awareness and understanding with users of this technology. For background information on the calculation of positional error (specifically, the calculation of RMSE or *root mean square error*), readers are referred to Section 5.2.2.

Errors related to the space segment

As a first source of error, the operators of the control segment may intentionally deteriorate radio signals of the satellites to the general public, to avoid optimal use of the system by the enemy, for instance in times of global political tension and war. This *selective availability*—meaning that the military forces allied with the control segment *will* still have access to undisturbed signals—may cause error that is an order of magnitude larger than all other error sources combined.⁹

Secondly, the satellite message may contain incorrect information. Assuming that it will always know its own identifier, the satellite may make two kinds of error:

1. *Incorrect clock reading:* Even atomic clocks can be off by a small margin, and since Einstein, we know that travelling clocks are slower than resident clocks, due to a so-called relativistic effect. If one understands that a clock that is off by 0.000001 sec causes an computation error in the satellite's pseudorange of approximately 300 m, it is clear that these satellite clocks require very strict monitoring.
2. *Incorrect orbit position:* The orbit of a satellite around our planet is easy to describe mathematically if both bodies are considered point masses, but in real life they are not. For the same reasons that the Geoid is not a simply shaped surface, the Earth's gravitation field that a satellite experiences

⁹Selective availability was stopped at the beginning of May 2000, and in late 2007 the White House decided to remove selective availability capabilities completely. However, the US government still has a range of capabilities and technology to implement regional denial of service of civilian GPS signals when needed in an area of conflict, effectively producing the same result.

in orbit is not simple either. Moreover, it is disturbed by solar and lunar gravitation, making its flight path slightly erratic and difficult to forecast exactly.

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

Errors related to the medium

Thirdly, the *medium* between sender and receiver may be of influence to the radio signals. The middle atmospheric layers of strato- and mesosphere are relatively harmless and of little hindrance to radio waves, but this is not true of the lower and upper layer. They are, respectively:

- *The troposphere*: the approximate 14 km high airspace just above the Earth's surface, which holds much of the atmosphere's oxygen and which envelopes all phenomena that we call the weather. It is an obstacle that delays radio waves in a rather variable way.
- *The ionosphere*: the most outward part of the atmosphere that starts at an altitude of 90 km, holding many electrically charged atoms, thereby forming a protection against various forms of radiation from space, including to some extent radio waves. The degree of ionization shows a distinct night and day rhythm, and also depends on solar activity.

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

Errors related to the receiver's environment

Fourth in this list is the error occurring when a radio signal is received via two or more paths between sender and receiver, some of which typically via a bounce off of some nearby surface, like a building or rock face. The term applied to this phenomenon is *multi-path*; when it occurs the multiple receptions of the same signal may interfere with each other (see Figure 4.26). Multi-path is a difficult to avoid error source.

Multi-path error

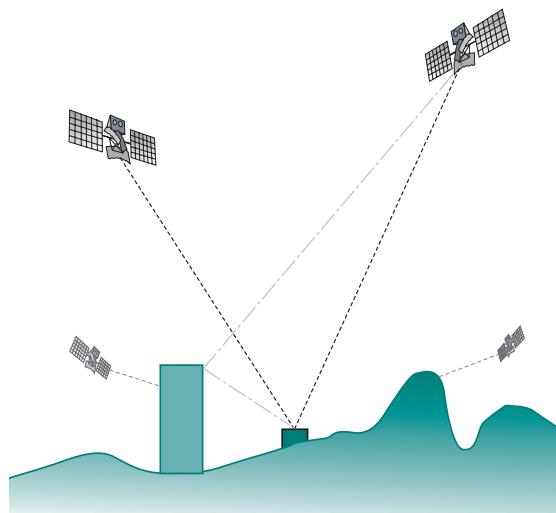


Figure 4.26: At any point in time, a number of satellites will be above the receiver's horizon. But not all of them will be 'in view' (like the left and right satellites), and for others multi-path signal reception may occur.

All of the above error sources have an influence on the computation of a satellite's pseudorange. In accumulation, they are called the *user equivalent range error* (UERE). Some error sources may be at work for all satellites being used by the receiver, for instance, selective availability and the atmospheric delay, while others

Range error

may be specific to one satellite, for instance, incorrect satellite information and multi-path.

Errors related to the relative geometry of satellites and receiver

There is one more source of error that is unrelated to individual radio signal characteristics, but that rather depends on the combination of the satellite signals used for positioning. Of importance is their constellation in the sky from the receiver perspective. Referring to Figure 4.27, one will understand that the sphere intersection technique of positioning will provide more precise results when the four satellites are nicely spread over the sky, and thus that the satellite constellation of Figure 4.27(b) is preferred over the one of 4.27(a). This error source is known as geometric dilution of precision (GDOP). GDOP is lower when satellites are just above the horizon in mutually opposed compass directions. However, such satellite positions have bad atmospheric delay characteristics, so in practice it is better if they are at least 15° above the horizon. When more than four satellites are in view, modern receivers use 'least-squares' adjustment to calculate the best positional fix possible from all of the signals. This gives a better solution than just using the "best four", as was done previously.

Geometric dilution of precision

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

Table 4.4: Indication of typical magnitude of errors in absolute satellite-based positioning

These errors are not all of similar magnitude. An overview of some typical val-

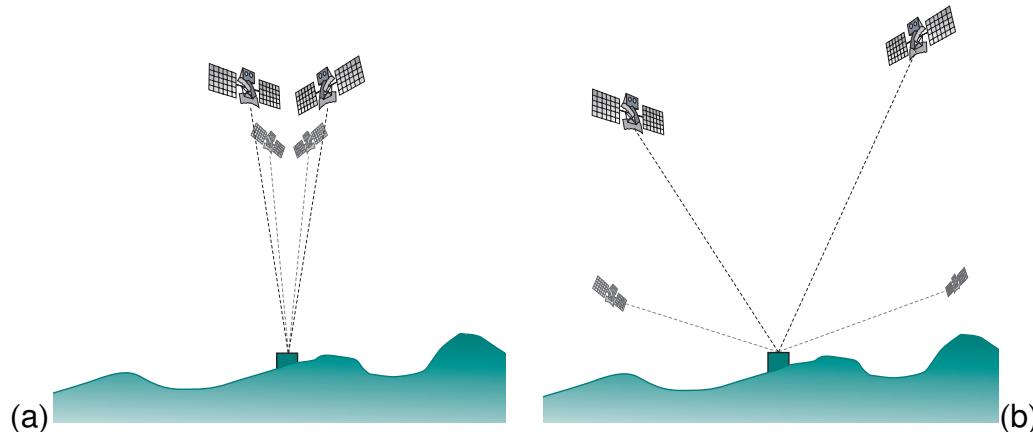


Figure 4.27: Geometric dilution of precision. The four satellites used for positioning can be in a bad constellation (a) or in a better constellation (b).

ues (without selective availability) is provided in Table 4.4. GDOP functions not so much as an independent error source but rather as a multiplying factor, decreasing the precision of position and time values obtained.

The procedure that we discussed above is known as *absolute, single-point positioning based on code measurement*. It is the fastest and simplest, yet least accurate way of determining a position using satellites. It suffices for recreational purposes and other applications that require horizontal accuracy not under 5–10 m. Typically, when encrypted military signals can also be used, on a dual-frequency receiver the achievable horizontal accuracy is 2–5 m. Below, we discuss other satellite-based positioning techniques with better accuracies.

4.2.3 Relative positioning

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

Random and systematic
error

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

In an optimal setting, reference and target receiver experience identical conditions and are connected by a direct data link, allowing the target to receive correctional data from the reference. In practice, relative positioning allows reference and target receiver to 70–200 km apart, and they will essentially experience similar atmospheric signal error. For each satellite in view, the reference receiver

¹⁰Please refer to section 5.2.2 for more detail on measurement error.

will determine its pseudorange error. After all, its position is known with high accuracy, so it can solve any pseudorange equations to determine the error. Subsequently, the target receiver, having received the error characteristics will apply the correction for each of the four satellite signals that it uses for positioning. In so doing, it can narrow down its accuracy to the 0.5–5 m range.

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

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

4.2.4 Network positioning

After discussing the advantages of relative positioning, we can move on to the notion of *network positioning*: an integrated, systematic network of reference receivers covering a large area like a continent or even the whole globe.

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

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

4.2.5 Code versus phase measurements

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

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

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

4.2.6 Positioning technology

We include this section to provide the reader with a little information on currently available satellite-based positioning technology. It should be noted that this textbook will easily outlive the currency of the information contained within it, as our technology is constantly evolving.

At present, two satellite-based positioning systems are operational (GPS and GLONASS), and a third is in the implementation phase (Galileo). Respectively, these are American, Russian and European systems. Any of these, but especially GPS and Galileo, will be improved over time, and will be augmented with new techniques.

GPS

The NAVSTAR Global Positioning System (GPS) was declared operational in 1994, providing Precise Positioning Services (PPS) to US and allied military forces as well as US government agencies, and Standard Positioning Services (SPS) to civilians throughout the world. Its space segment nominally consists of 24 satellites, each of which orbit our planet in 11h58m at an altitude of 20,200 km. There can be any number of satellites active, typically between 21 and 27. The satellites are organized in six orbital planes, somewhat irregularly spaced, with an angle of inclination of 55–63° with the equatorial plane, nominally having four satellites each (see Figure 4.28). This means that a receiver on Earth will have between five and eight (sometimes up to twelve) satellites in view at any point in time. Software packages exist to help in planning GPS surveys, identifying expected satellite set-up for any location and time.

Orbital planes

GPS's control segment has its master control in Colorado, US, and monitor stations in a belt around the equator, namely in Hawaii, Kwajalein Atoll in the Marshall Islands, Diego Garcia (British Indian Ocean Territory) and Ascension Island (UK, southern Atlantic Ocean).

The NAVSTAR satellites transmit two radio signals, namely the L1 frequency at 1575.42 MHz and the L2 frequency at 1227.60 MHz. There are also a third and fourth signal, but they are not important for our discussion here. The first two signals consist of:

- The carrier waves at the given frequencies,
- A coarse ranging code, known as C/A, modulated on L1,

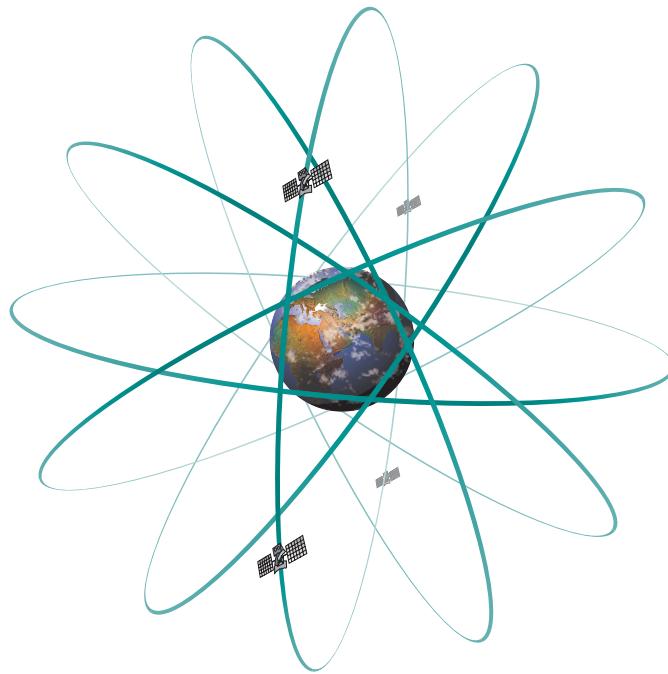


Figure 4.28: Constellation of satellites, four shown in only one orbit plane, in the GPS system.

- An encrypted precision ranging code, known as $P(Y)$, modulated on L1 and L2, and
- A navigation message modulated on both L1 and L2.

The role of L2 is to provide a second radio signal, thereby allowing (the more expensive) dual-frequency receivers a way of determining fairly precisely the actual ionospheric delay on satellite signals received. The role of the ranging codes is two-fold:

1. To identify the satellite that sent the signal, as each satellite sends unique codes, and the receiver has a look-up table for these codes, and
2. To determine the signal transit time, and thus the satellite's pseudorange.

The navigation message contains the satellite orbit and satellite clock error information, as well as some general system information. GPS also carries a fifth, encrypted military signal carrying the M-code. GPS uses WGS84 as its reference system. It has been refined on several occasions and is now aligned with the ITRF at the level of a few centimetres worldwide. (See also Section 4.1.1.) GPS has adopted UTC as its time system.

WGS84 and ITRF

In the civil market, GPS receivers of varying quality are available, their quality depending on the embedded positioning features: supporting single- or dual-frequency, supporting only absolute or also relative positioning, performing code measurements or also carrier phase measurements. Leica and Trimble are two of the well-known brands in the high-precision, professional surveying domain; Magellan and Garmin, for instance, operate in the lower price, higher volume consumer market range, amongst others for recreational use in outdoor activities. Many of these are single frequency receivers, doing only code measurements, though some are capable of relative positioning. This includes the new generation of GPS-enabled mobile phones.

GPS manufacturers

GLONASS

What GPS is to the US military, is GLONASS to the Russian military, specifically the Russian Space Forces. Both systems were primarily designed on the basis of military requirements. The big difference between the two is that GPS generated a major interest in civil applications, thus having an important economic impact. This cannot be said of GLONASS.

The GLONASS space segment consists of nominally 24 satellites, organized in three orbital planes, with an inclination of 64.8° with the equator. Orbiting altitude is 19,130 km, with a period of revolution of 11 hours 16 min. GLONASS uses the PZ-90 as its reference system, and like GPS uses UTC as time reference, though with an offset for Russian daylight.

GLONASS radio signals are somewhat similar to that of GPS, but differ in the details. Satellites use different identifier schemes, and their navigation message use other parameters. They also use different frequencies: GLONASS L1 is at approximately 1605 MHz (changes are underway), and L2 is at approximately 1248 MHz. Otherwise, the GLONASS system performance is rather comparable to that of GPS.

Galileo

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

Galileo is the name of this EU system. The vision is that satellite-based positioning will become even bigger due to the emergence of mobile phones equipped with receivers, perhaps with some 400 million users by the year 2015. Development of the system has experienced substantial delays, and at the time of writing European ministers insist that Galileo should be up and running by the end of 2013. The completed system will have 27 satellites, with three in reserve, orbiting in one of three, equally spaced, circular orbits at an elevation of 23,222 km, inclined 56° with the equator. This higher inclination, when compared to that of GPS, has been chosen to provide better positioning coverage at high latitudes, such as northern Scandinavia where GPS performs rather poorly.

In June 2004, the EU and the US agreed to make Galileo and GPS compatible by adoption of interchangeable satellite signal set-ups. The effect of this agreement is that the Galileo/GPS tandem satellite system will have so many satellites in the sky (close to 60) that a receiver can almost always find an optimal constellation in view. This will be especially useful in situations where in the past bad signal reception happened: in built-up areas and forests, for instance. It will also bring the implementation of a Global Navigation Satellite System (GNSS) closer as positional accuracy and reliability will improve. With such a system, eventually one expects to implement fully automated air and road traffic. Automatic aircraft landing, for instance, requires horizontal accuracy in the order of

4 m, and vertical accuracy below 1 m: these requirements can currently not be achieved reliably.

The Galileo Terrestrial Reference Frame (GTRF) will be a realization of the ITRS independently set up from that of GPS, so that one system can back-up for the other. Positional differences between the WGS84 and the GTRF will be at worst a few centimetres. The Galileo System Time (GST) will closely follow International Atomic Time (TAI) with a time offset of less than 50 nsec for 95 % of the time over any period of a year. Information on the actual offset between GST and TAI, and between GST and UTC (as used in GPS) will be broadcasted in the Galileo satellite signal.

TAI

Satellite-based augmentation systems

Satellite-based augmentation systems (SBAS) aim to improve accuracy and reliability of satellite-based positioning (see the section on network positioning, page 256) in support of safety-critical navigation applications such as aircraft operations near airfields. The typical technique is to provide an extra, now geostationary, satellite that has a large service area like a continent, and which sends differential data about standard positioning satellites that are currently in view in its service area. If multiple ground reference stations are used, the quality of the differential data can be quite good and reliable. Signals typically use the frequency already in use by the positioning satellites, so that receivers can receive the differential code without problem.

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

Currently, three systems are operational: for North America WAAS (Wide-Area Augmentation System) is in place, EGNOS (European Geostationary Navigation Overlay Service) for Europe, and MSAS (Multi-functional Satellite Augmentation System) for eastern Asia. The ground segment of WAAS consists of 24 con-

trol stations, spread over North America; that of EGNOS has 34 stations. These three systems are compatible, guaranteeing international coverage.

Signals of the respective satellites (under various names like AOR, Artemis, IOR, Inmarsat, MTSAT) can usually be received outside their respective service areas, but the use of these signals is discouraged, as they will not help improve positional accuracy. Satellite identifiers, as shown in the receiver, have numbers above 30, setting them apart from standard positioning satellites.

Summary

This chapter focuses upon locating objects and events on the Earth's surface. In this context, a number of principles related to spatial reference systems, including vertical and horizontal datums were discussed in Section 4.1.¹¹

To summarise, each projection and datum has particular characteristics that make it useful for specific mapping purposes. A projection is chosen to minimize the errors for the area and relevant to the scale of the mapping project being undertaken, and the required distortion property, which in turn depends on the purpose for which the map will be used. We need to be aware of issues brought about by the combination of spatial data from different sources that use different reference systems. This issue is becoming increasingly important, as more and more data is being shared. Often, transformations are necessary to enable the combination of disparate data layers.

Section 4.2 discussed the various methods of satellite-based positioning, from basic principles to characteristics of current implementations, and the different levels of accuracy associated with each of these methods. This included a discussion of sources of error in the context of both absolute and relative positioning. Key aspects of positional accuracy are dealt with in more detail in the following chapter, in the context of data quality.

¹¹This section is accompanied by a website at <http://kartoweb.itc.nl/geometrics>. Here, interested readers can find more background information and a list of frequently asked questions.

Questions

1. 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?
2. On page [196](#) we mentioned that in geodetic practice the definition of an ellipsoid is usually by its semi-major axis a and flattening f . Flattening is dependent on both the semi-major axis a and the semi-minor axis b . Assume that the semi-major axis a of an ellipsoid is 6378137 m and the flattening f is 1:298.257. Using these facts determine the semi-minor axis b (make use of the given equations).
3. You are required to match GPS data with some map data. The GPS data and the map layer are based on different horizontal datums. Which steps should you take to make the GPS data spatially compatible with the map data?
4. Suppose you wish to produce a small scale thematic map of your country. The map should show the population densities for the different regions (or provinces). What would be a good map projection for the representation of the population densities of your country? Consider the class of the projection, the projection property and the line(s) of intersection or the point or line of tangency.

5. In section 4.2.1, we discussed the principles of absolute positioning. To a large extent, there is an analogy with how human beings assess the risks of a thunder storm with lightning. Explain that analogy, and discuss the 'measuring errors'.
6. Estimate a realistic distance between a GPS satellite that is in view and the receiver that you are holding, clarifying your assumptions. Indicate minimum and maximum values. Finally, compute the time delay a satellite message incurs before being received, again clarifying the assumptions made.
7. On page 247 we mentioned size of the pseudorange error with respect to satellite clock error. Think up why the estimates were as given. Also analyse, using geometric arguments, what positioning error might result from a single satellite clock error of 0.000001 sec.
8. How could one force a GPS receiver to operate in 2D positioning mode? How would one set up an experiment to determine positioning accuracy in this mode and the relation to actual height of the receiver?

Chapter 5

Data entry and preparation

Spatial data can be obtained from various sources. It can be collected from scratch, using direct spatial data acquisition techniques, or indirectly, by making use of existing spatial data collected by others. Under the first heading we could include field survey data and remotely sensed images. Under the second fall paper maps and existing digital data sets.

This chapter discusses the collection and use of data under both of these headings. It seeks to prepare users of spatial data by drawing attention to issues concerning data accuracy and quality. A range of procedures for data checking and clean-up are discussed to prepare data for analysis, including several methods for interpolating point data.

5.1 Spatial data input

5.1.1 Direct spatial data capture

One way to obtain spatial data is by *direct observation* of the relevant geographic phenomena. This can be done through ground-based field surveys, or by using remote sensors in satellites or airplanes (see Chapter 4). Many Earth sciences have developed their own survey techniques, as ground-based techniques remain the most important source for reliable data in many cases.

Primary data

Data which is captured directly from the environment is known as *primary data*.

With primary data the core concern in knowing its properties is to know the process by which it was captured, the parameters of any instruments used and the rigour with which quality requirements were observed.

Remotely sensed imagery is usually not fit for immediate use, as various sources of error and distortion may have been present, and the imagery should first be freed from these. This is the domain of remote sensing, and these issues are discussed further in *Principles of Remote Sensing* [53]. In the context of this book,

An *image* refers to raw data produced by an electronic sensor, which are not pictorial, but arrays of digital numbers related to some property of an object or scene, such as the amount of reflected light.

For an image, no interpretation of reflectance values as thematic or geographic characteristics has taken place. When the reflectance values have been translated into some ‘thematic’ variable, we refer to it as a raster. Section 2.3.1 provides

Images and rasters

more detail on rasters. It is interesting to note that we refer to image *pixels* but to raster *cells*, although both are stored in a GIS in the same way.

In practice, it is not always feasible to obtain spatial data by direct spatial data capture. Factors of cost and available time may be a hindrance, or previous projects sometimes have acquired data that may fit the current project's purpose.

5.1.2 Indirect spatial data capture

In contrast to direct methods of data capture described above, spatial data can also be sourced *indirectly*. This includes data derived from existing paper maps through scanning, data digitized from a satellite image, processed data purchased from data capture firms or international agencies, and so on. This type of data is known as *secondary data*:

Secondary data

Any data which is not captured directly from the environment is known as *secondary data*.

Below we discuss key sources of secondary data and issues related to their use in analysis of which the user should be aware.

Digitizing

A traditional method of obtaining spatial data is through *digitizing* existing paper maps. This can be done using various techniques. Before adopting this approach, one must be aware that positional errors already in the paper map will further accumulate, and one must be willing to accept these errors.

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 surface (the tablet), while in on-screen digitizing, a scanned image of the map (or some other image) is shown on the computer screen. In both of these forms, an operator follows the map's features (mostly lines) with a mouse device, thereby tracing the lines, and storing location coordinates relative to a number of previously defined *control points*. The function of these points is to 'lock' a coordinate system onto the digitized data: the control points on the map have *known* coordinates, 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.

Control points

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.

Scanning

An ‘office’ scanner illuminates a document and measures the intensity of the reflected light with a CCD array. The result is an image as a matrix of pixels, each of which holds an intensity value. Office scanners have a fixed maximum resolution, expressed as the highest number of pixels they can identify per inch; the unit is dots-per-inch (dpi). For manual on-screen digitizing of a 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.

Resolution

(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 image processing techniques. It is important to understand that scanning does *not* result in a structured data set of classified and coded objects. Additional work is required to recognize features and to associate categories and other thematic attributes with them.

Vectorization

The process of distilling points, lines and polygons from a scanned image is called *vectorization*. As scanned lines may be several pixels wide, they are often first thinned to retain only the centreline. The remaining centreline pixels are converted to series of (x, y) coordinate pairs, defining a polyline. Subsequently, features are formed and attributes are attached to them. This process may be entirely automated or performed semi-automatically, with the assistance of an operator. Pattern recognition methods—like Optical Character Recognition (OCR) for text—can be used for the automatic detection of graphic symbols and text.

OCR

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 5.1.

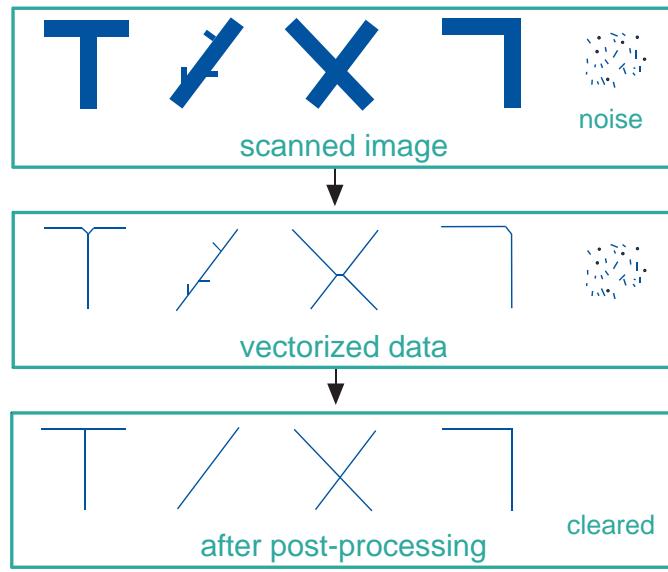


Figure 5.1: The phases of the vectorization process and the various sorts of small error caused by it. The post-processing phase makes the final repairs.

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.

In practice, the optimal choice may be a combination of methods. For example, contour line film separations can be automatically digitized and used to produce a DEM. Existing topographic maps must be digitized manually, but new, geometrically corrected aerial photographs, with vector data from the topographic maps displayed directly over it, can be used for updating existing data files by means of manual on-screen digitizing.

5.1.3 Obtaining spatial data elsewhere

Over the past two decades, spatial data has been collected in digital form at increasing rate, stored in various databases by the individual producers for their own use and for commercial purposes. More and more of this data is being shared among GIS users. This is for several reasons. Some of this data is freely available, although other data is only available commercially, as is the case for most satellite imagery. High quality data remain both costly and time-consuming to collect and verify, as well as the fact that more and more GIS applications are looking at not just local, but national or even global processes. As we will see below, new technologies have played a key role in the increasing availability of geospatial data. As a result of this increasing availability, we have to be more careful that the data we have acquired is of sufficient *quality* to be used in analysis and decision making. For this reason, we discuss key data quality parameters in Section 5.2.

Clearinghouses and web portals

Spatial data can also be acquired from centralized repositories. More often those repositories are embedded in Spatial Data Infrastructures (see Section 3.2.3), which make the data available through what is sometimes called a spatial data *clearinghouse*. This is essentially a marketplace where data users can ‘shop’. It will be no surprise that such markets for digital data have an entrance through the world wide web. The first entrance is typically formed by a web portal which categorizes all available data and provides a local search engine and links to data documentation (also called metadata). It often also points to data viewing and processing services. Standards-based geo-webservices have become the common technology behind such portal services (see below for further detail).

Spatial Data Infrastructures

Metadata

Metadata is defined as background information that describes all necessary information about the data itself. More generally, it is known as ‘data about data’. This includes:

- *Identification information*: Data source(s), time of acquisition, *etc.*
- *Data quality information*: Positional, attribute and temporal accuracy, lineage, *etc.*
- *Entity and attribute information*: Related attributes, units of measure, *etc.*

In essence, metadata answer *who*, *what*, *when*, *where*, *why*, and *how* questions about all facets of the data made available. Maintaining metadata is an key part in maintaining data and information quality in GIS. This is because it can serve different purposes, from description of the data itself through to providing instructions for data handling. Depending on the type and amount of metadata provided, it could be used to determine the data sets that exist for a geographic location, evaluate whether a given data set meets a specified need, or to process and use a data set.

Data formats and standards

An important problem in any environment involved in digital data exchange is that of *data formats* and *data standards*. Different formats were implemented by different GIS vendors; different standards came about with different standardization committees. The phrase 'data standard' refers to an agreed upon way of representing data in a system in terms of content, type and format. The good news about both formats and standards is that there are many to choose from; the bad news is that this can lead to a range of conversion problems. Several metadata standards standards for digital spatial data exist, including the International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC) standards.

ISO and OGC standards

5.2 Data quality

With the advent of satellite remote sensing, GPS and GIS technology, and the increasing availability of digital spatial data, resource managers and others who formerly relied on the surveying and mapping profession to supply high quality map products are now in a position to produce maps themselves. At the same time, GISs are being increasingly used for *decision support* applications, with increasing reliance on secondary data sourced through data providers or via the internet, through geo-webservices. The implications of using low-quality data in important decisions are potentially severe. There is also a danger that uninformed GIS users introduce errors by incorrectly applying geometric and other transformations to the spatial data held in their database.

Application requirements

Below we look at the main issues related to data quality in spatial data. As outlined in Section 1.1.4, we will discuss positional, temporal and attribute accuracy, lineage, completeness, and logical consistency. We will begin with a brief discussion of the terms accuracy and precision, as these are often taken to mean the same thing. For a more detailed discussion and advanced topics relating to data quality, the reader is referred to [17].

5.2.1 Accuracy and precision

So far we have used the terms error, accuracy and precision without appropriately defining them. Accuracy should not be confused with *precision*, which is a statement of the smallest unit of measurement to which data can be recorded. In conventional surveying and mapping practice, accuracy and precision are closely related. Instruments with an appropriate precision are employed, and surveying methods chosen, to meet specified 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 m² from coordinates obtained by digitizing a 1 : 50,000 map.

Accuracy tolerances

Using graphs that display the probability distribution (for which see below) of a measurement against the true value T , the relationship between accuracy and precision can be clarified. In Figure 5.2, we depict the cases of good/bad accuracy against good/bad precision.¹ An *accurate* measurement has a mean close to the true value; a *precise* measurement has a sufficiently small variance.

¹Here we use the terms ‘good’ and ‘bad’ to illustrate the extremes of both accuracy and precision. In real world terms we refer to whether data is ‘fit for use’ for a given application.

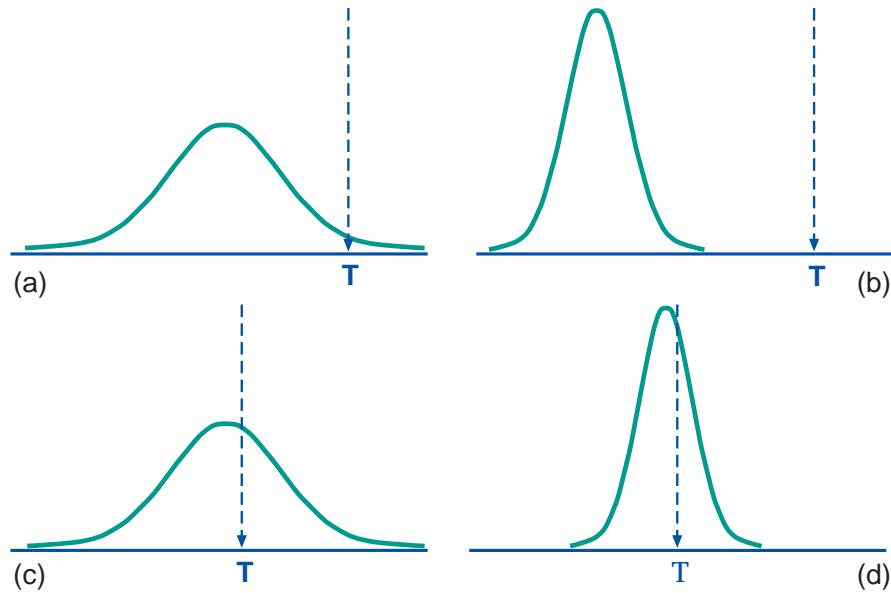


Figure 5.2: A measurement probability function and the underlying true value T : (a) bad accuracy and precision, (b) bad accuracy/good precision, (c) good accuracy/bad precision, and (d) good accuracy and precision.

5.2.2 Positional accuracy

The surveying and mapping profession has a long tradition of determining and minimizing errors. This applies particularly to land surveying and photogrammetry, both of which tend to regard positional and height errors as undesirable. Cartographers also strive to reduce geometric and attribute errors in their products, and, in addition, define quality in specifically cartographic terms, for example quality of linework, layout, and clarity of text.

It must be stressed that all measurements made with surveying and photogrammetric instruments are subject to error. These include:

1. Human errors in measurement (e.g. reading errors) generally referred to as gross errors or *blunders*. These are usually large errors resulting from carelessness which could be avoided through careful observation, although it is never absolutely certain that all blunders have been avoided or eliminated.
2. Instrumental or *systematic* errors (e.g. due to misadjustment of instruments). This leads to errors that vary systematically in sign and/or magnitude, but can go undetected by repeating the measurement with the same instrument. Systematic errors are particularly dangerous because they tend to accumulate.
3. So-called *random* errors caused by natural variations in the quantity being measured. These are effectively the errors that remain after blunders and systematic errors have been removed. They are usually small, and dealt with in least-squares adjustment.

Error sources

Section 4.2 discussed the errors inherent in various methods of spatial positioning. Below we will look at more general ways of quantifying positional accuracy using *root mean square error (RMSE)*.

Measurement errors are generally described in terms of *accuracy*. In the case of spatial data, accuracy may relate not only to the determination of coordinates (positional error) but also to the measurement of quantitative attribute data. The accuracy of a single measurement can be defined as:

"the closeness of observations, computations or estimates to the true values or the values perceived to be true" [41].

In the case of surveying and mapping, the 'truth' is usually taken to be a value obtained from a survey of higher accuracy, for example by comparing photogrammetric measurements with the coordinates and heights of a number of independent check points determined by field survey. Although it is useful for assessing the quality of definite objects, such as cadastral boundaries, this definition clearly has practical difficulties in the case of natural resource mapping where the 'truth' itself is uncertain, or boundaries of phenomena become fuzzy. This type of uncertainty in natural resource data is elaborated upon on page 295.

Prior to the availability of GPS, resource surveyors working in remote areas sometimes had to be content with ensuring an acceptable degree of *relative accuracy* among the measured positions of points within the surveyed area. If location and elevation are fixed with reference to a network of control points that are assumed to be free of error, then the *absolute accuracy* of the survey can be determined.

Relative and absolute accuracy

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 (coordinate values from an independent source of higher accuracy for identical points). The differences at each point can be plotted as error vectors, as is done in Figure 5.3 for a single measurement. The error vector can be seen as having constituents in the x - and y -directions, which can be recombined by vector addition to give the error vector representing its locational error.

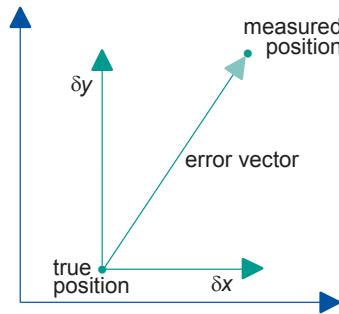


Figure 5.3: The positional error of a measurement can be expressed as a vector, which in turn can be viewed as the vector addition of its constituents in x - and y -direction, respectively δx and δy .

For each checkpoint, the error vector has components δx and δy . The observed errors should be checked for a *systematic* error component, which may indicate a (possibly repairable) lapse in the measurement method. Systematic error has occurred when $\sum \delta x \neq 0$ or $\sum \delta y \neq 0$.

The systematic error $\delta \bar{x}$ in x is then defined as the average deviation from the

true value:

$$\delta\bar{x} = \frac{1}{n} \sum_{i=1}^n \delta x_i.$$

Analogously to the calculation of the variance and standard deviation of a statistical sample, the root mean square errors m_x and m_y of a series of coordinate measurements are calculated as the square root of the average squared deviations:

$$m_x = \sqrt{\frac{1}{n} \sum_{i=1}^n \delta x_i^2} \quad \text{and} \quad m_y = \sqrt{\frac{1}{n} \sum_{i=1}^n \delta y_i^2},$$

where δx^2 stands for $\delta x \cdot \delta x$. The total RMSE is obtained with the formula

$$m_{\text{total}} = \sqrt{m_x^2 + m_y^2},$$

which, by the Pythagorean rule, is the length of the average (root squared) vector.

Accuracy tolerances

Many kinds of measurement can be naturally represented by a bell-shaped probability density function p , as depicted in Figure 5.4(a). This function is known as the *normal (or Gaussian) distribution* of a continuous, random variable, in the figure indicated as Y . Its shape is determined by two parameters: μ , which is the mean expected value for Y , and σ which is the standard deviation of Y . A small σ leads to a more attenuated bell shape.

Distribution of errors

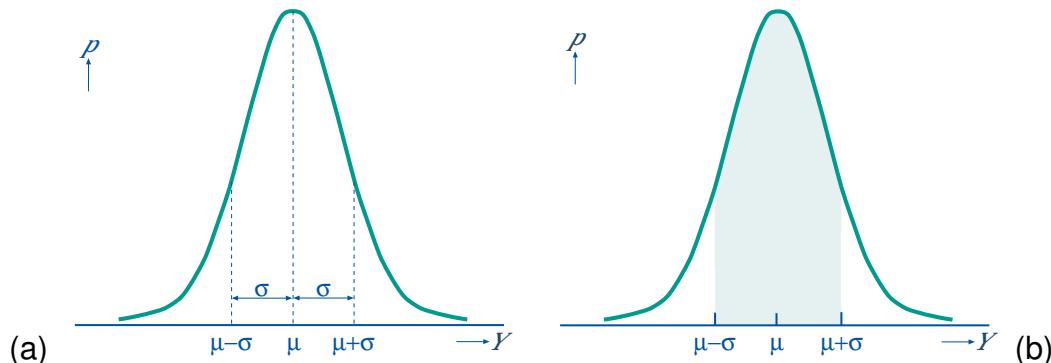


Figure 5.4: (a) Probability density function p of a variable Y , with its mean μ and standard deviation σ .
(b) The probability that Y is in the range $[\mu - \sigma, \mu + \sigma]$.

Any probability density function p has the characteristic that the area between its curve and the horizontal axis has size 1. Probabilities P can be inferred from p as the size of an area under p 's curve. Figure 5.4(b), for instance, depicts $P(\mu - \sigma \leq Y \leq \mu + \sigma)$, i.e. the probability that the value for Y is within distance σ from μ . In a normal distribution this specific probability for Y is always 0.6826.

The *RMSE* can be used to assess the probability that a particular set of measurements does not deviate too much from, i.e. is within a certain range of, the 'true' value. In the case of coordinates, the probability density function often is

considered to be that of a two-dimensional normally distributed variable (see Figure 5.5). The three standard probability values associated with this distribution are:

- 0.50 for a circle with a radius of $1.1774 m_x$ around the mean (known as the *circular error probable*, CEP);
- 0.6321 for a circle with a radius of $1.412 m_x$ around the mean (known as the *root mean square error*, RMSE);
- 0.90 for a circle with a radius of $2.146 m_x$ around the mean (known as the *circular map accuracy standard*, CMAS).

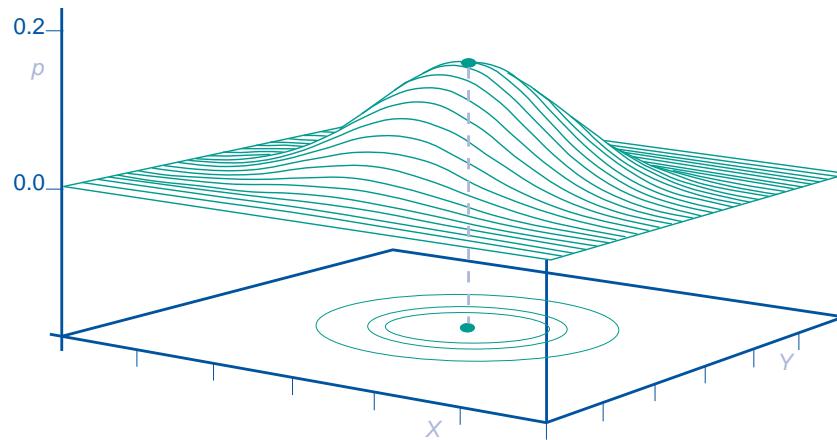


Figure 5.5: Probability density p of a normally distributed, two-dimensional variable (X, Y) (also known as a normal, bivariate distribution). In the ground plane, from inside out, are indicated the circles respectively associated with CEP, RMSE and CMAS.

The RMSE provides an estimate of the spread of a series of measurements around their (assumed) ‘true’ values. It is therefore commonly used to assess the quality

of transformations such as the absolute orientation of photogrammetric models or the spatial referencing of satellite imagery. The RMSE also forms the basis of various statements for reporting and verifying compliance with defined map accuracy *tolerances*. An example is the American National Map Accuracy Standard, which states that:

"No more than 10% of well-defined points on maps of 1:20,000 scale or greater may be in error by more than 1/30 inch."

Normally, compliance to this tolerance is based on at least 20 well-defined check-points.

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 5.6). 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 5.7).

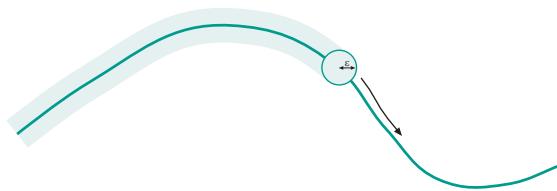


Figure 5.6: The ε - or Perkal band is formed by rolling an imaginary circle of a given radius along a line.

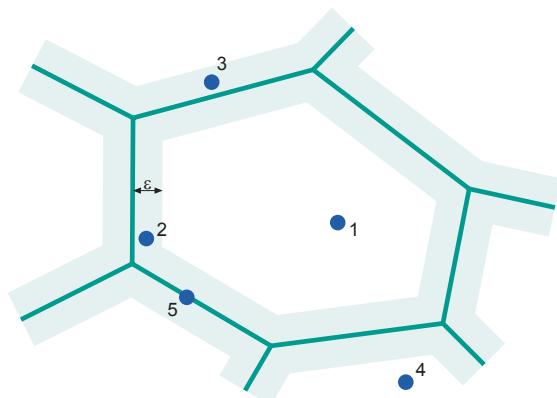


Figure 5.7: The ε -band may be used to assess the likelihood that a point falls within a particular polygon. Source: [43]. Point 3 is less likely part of the middle polygon than point 2.

Describing natural uncertainty in spatial data

There are many situations, particularly in surveys of natural resources, where, according to Burrough, “practical scientists, faced with the problem of dividing up undividable complex continua have often imposed their own crisp structures on the raw data” [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.

Classification

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 positional accuracy, such as RMSE (Figure 5.3), may be virtually insignificant in comparison to the uncertainty inherent in vegetation and soil boundaries, for example.

Boundaries

In conventional applications of the error matrix to assess the quality of nominal (categorical) data such as land use, individual samples can be considered in terms of Boolean set theory. The Boolean *membership function* is binary, i.e. an element is either 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.”

Membership functions

Since its original development by Zadeh [58], there has been considerable discussion of fuzzy, or continuous, set theory as an approach for handling imprecise spatial data. In GIS, fuzzy set theory appears to have two particular benefits:

Fuzzy set theory

1. The ability to handle logical modelling (map overlay) operations on inexact data, and
2. The possibility of using a variety of natural language expressions to qualify uncertainty.

Unlike Boolean sets, fuzzy or continuous sets have a membership function, which can assign to a member any value between 0 and 1 (see Figure 5.8). The membership function of the Boolean set of Figure 5.8(a) can be defined as MF^B follows:

$$MF^B(x) = \begin{cases} 1 & \text{if } b_1 \leq x \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$

The crisp and uncertain set membership functions of Figure 5.8 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 5.8(b), in contrast to function

MF^B above, can be defined as a function MF^C , following Heuvelink in [21]:

$$\text{MF}^C(x) = \begin{cases} \frac{1}{1 + \left(\frac{x - b_1}{d_1}\right)^2} & \text{if } x < b_1 \\ 1 & \text{if } b_1 \leq x \leq b_2 \\ \frac{1}{1 + \left(\frac{x - b_2}{d_2}\right)^2} & \text{if } x > b_2 \end{cases}$$

The parameters d_1 and d_2 denote the width of the transition zone around the kernel of the class such that $\text{MF}^C(x) = 0.5$ at the thresholds $b_1 - d_1$ and $b_2 + d_2$, respectively. If d_1 and d_2 are both zero, the function MF^C reduces to MF^B .

An advantage of fuzzy set theory is that it permits the use of natural language to describe uncertainty, for example, “near,” “east of” and “about 23 km from,” as such natural language expressions can be more faithfully represented by appropriately chosen membership functions.

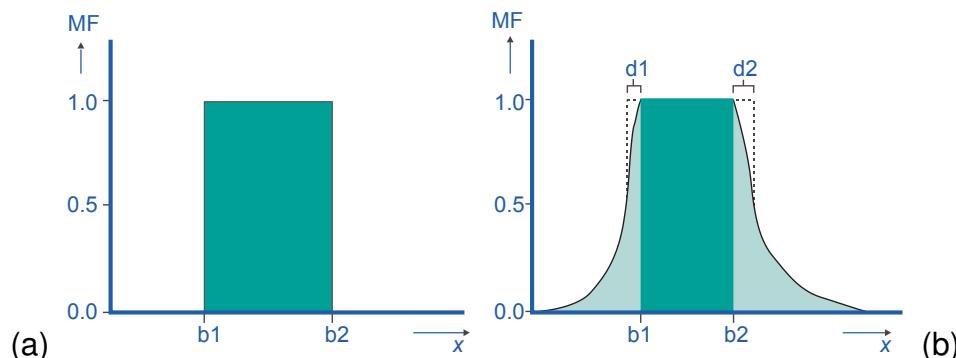


Figure 5.8: (a) Crisp (Boolean) and (b) uncertain (fuzzy) membership functions MF. After Heuvelink [21]

5.2.3 Attribute accuracy

We can identify two types of attribute accuracies. These relate to the type of data we are dealing with:

- For *nominal or categorical* data, the accuracy of labeling (for example the type of land cover, road surface, etc).
- For *numerical* data, numerical accuracy (such as the concentration of pollutants in the soil, height of trees in forests, etc).

It follows that depending on the data type, assessment of attribute accuracy may range from a simple check on the labelling of features—for example, is a road classified as a metalled road actually surfaced or not?—to complex statistical procedures for assessing the accuracy of numerical data, such as the percentage of pollutants present in 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 (also known as a confusion or misclassification matrix) that can be used to evaluate the accuracy of the classification. An example is provided in Table 5.1, 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

Error matrix

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.

Classified image	Reference data			<i>total</i>
	<i>Forest</i>	<i>Agriculture</i>	<i>Urban</i>	
<i>Forest</i>	62	5	0	67
<i>Agriculture</i>	2	18	0	20
<i>Urban</i>	0	1	12	13
<i>total</i>	64	24	12	100

Table 5.1: Example of a simple error matrix for assessing map attribute accuracy. The overall accuracy is $(62+18+12)/100 = 92\%$.

For more details on attribute accuracy, the student is referred to *Principles of Remote Sensing* [53].

5.2.4 Temporal accuracy

As noted, the amount of spatial data sets and archived remotely sensed data has increased enormously over the last decade. These data can provide useful temporal information such as changes in land ownership and the monitoring of environmental processes such as deforestation. Analogous to its positional and attribute components, the quality of spatial data may also be assessed in terms of its *temporal accuracy*. For a static feature this refers to the difference in the values of its coordinates at two different times.

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 may change more frequently.

Consistency and validity

5.2.5 Lineage

Lineage describes the history of a data set. In the case of published maps, some lineage information may be provided as part of the metadata, in the form of a note on the data sources and procedures used in the compilation of the data. Examples include the date and scale of aerial photography, and the date of field verification. Especially for digital data sets, however, lineage may be defined more formally as:

“that part of the data quality statement that contains information that describes the source of observations or materials, data acquisition and compilation methods, conversions, transformations, analyses and derivations that the data has been subjected to, and the assumptions and criteria applied at any stage of its life.” [14]

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’.

5.2.6 Completeness

Completeness refers to whether there are data lacking in the database compared to what exists in the real world. Essentially, it is important to be able to assess what does and what does not belong to a *complete* dataset as intended by its producer. It might be incomplete (i.e. it is ‘missing’ features which exist in the real world), or overcomplete (i.e. it contains ‘extra’ features which do not belong within the scope of the data set as it is defined).

Incomplete and
overcomplete

Completeness can relate to either spatial, temporal, or thematic aspects of a data set. For example, a data set of property boundaries might be spatially incomplete because it contains only 10 out of 12 suburbs; it might be temporally incomplete because it does not include recently subdivided properties; and it might be thematically overcomplete because it also includes building footprints.

5.2.7 Logical consistency

For any particular application, (predefined) logical rules concern:

- The *compatibility* of data with other data in a data set (e.g. in terms of data format),
- The absence of any *contradictions* within a data set,
- The *topological consistency* of the data set, and
- The allowed attribute *value ranges*, as well as combinations of attributes. For example, attribute values for population, area, and population density must agree for all entities in the database.

The absence of any inconsistencies does not necessarily imply that the data are accurate.

5.3 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 require conversion to either vector format or raster format to match other data sets which will be used in the analysis. Additionally, the data preparation process includes associating attribute data with the spatial features through either manual input or reading digital attribute files into the GIS/DBMS.

The intended use of the acquired spatial data may require only a subset of the original data set, as only some of the features are relevant for subsequent analysis or subsequent map production. In these cases, data and/or cartographic generalization can be performed on the original data set.

Intended use

5.3.1 Data checks and repairs

Acquired data sets must be checked for quality in terms of the accuracy, consistency and completeness parameters discussed above. Often, errors can be identified automatically, after which manual editing methods can be applied to correct the errors. Alternatively, some software may identify and automatically correct certain types of errors. Below, we focus on the *geometric*, *topological*, and *attribute* components of spatial data.

Automatic and manual checking

'Clean-up' operations are often performed in a standard sequence. For example, crossing lines are split before dangling lines are erased, and nodes are created at intersections before polygons are generated. These are illustrated in Table 5.2.

With polygon data, one usually starts with many polylines, in an unwieldy format known as *spaghetti data*, that are combined in the first step (from Figure 5.9(a) to (b)). This results in fewer polylines with more internal vertices. Then, polygons can be identified (c). Sometimes, polylines that should connect to form closed boundaries do not, and therefore must be connected (either manually or automatically); this step is not indicated in the figure. In a final step, the elementary topology of the polygons can be derived (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 5.2: Clean-up operations for vector data

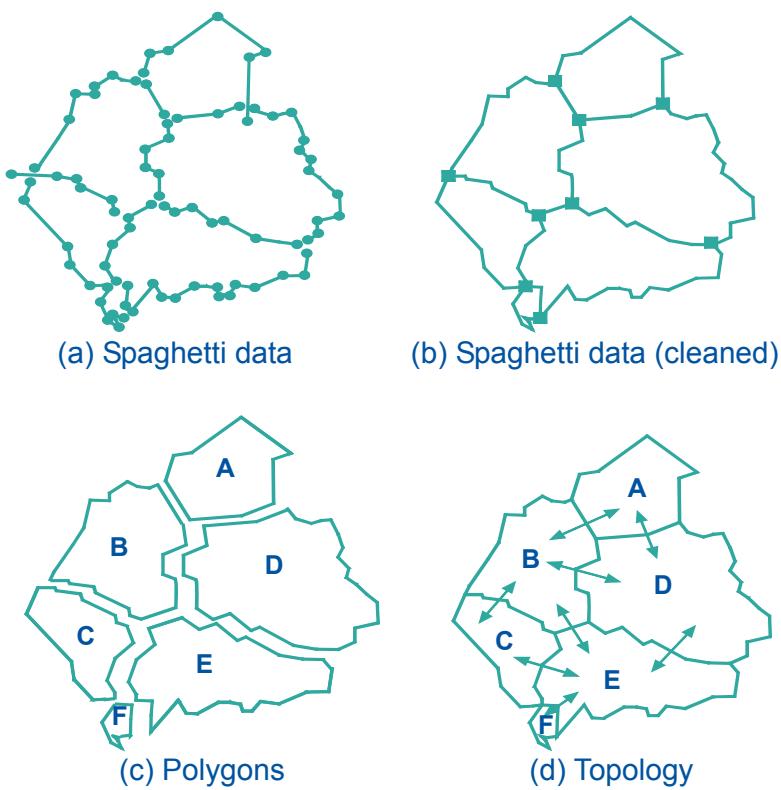


Figure 5.9: Successive clean-up operations for vector data, turning spaghetti data into topological structure.

Associating attributes

Attributes may be automatically associated with features that have unique identifiers. We have already discussed these techniques in Section 3.5. In the case of vector data, attributes are assigned directly to the features, while in a raster the attributes are assigned to all cells that represent a feature. Section 5.2.3 discusses issues relating to raster attribute accuracy in more detail.

Rasterization or vectorization

Vectorization 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*, which are dealt with in *Principles of Remote Sensing* [53].

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 cell size which is too large may result in cells that cover parts of multiple vector features, and then ambiguity arises as to what value to assign to the cell. If, on the other hand, the cell size is too small, the file size of the raster may increase significantly.

Rasterization itself could be seen as a ‘backwards step’: firstly, raster boundaries are only an *approximation* of the objects’ original boundary. Secondly, the original ‘objects’ can no longer be treated as such, as they have lost their topological properties. Often the reason for rasterisation is because it facilitates easier combination with other data sources also in raster formats, and/or because there are several analytical techniques which are easier to perform upon raster data (please refer to Chapter 6). An alternative 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.

Topology generation

We have already discussed derivation of topology from vectorized data sources. However, more topological relations may sometimes be needed, for instance in networks, e.g. the questions of line connectivity, flow direction, 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.

What kind of topology is required?

5.3.2 Combining data from multiple sources

A GIS project usually involves multiple data sets, so the next step addresses the issue of how these multiple sets relate to each other. There are four fundamental cases to be considered in the combination of data from different sources:

1. They may be about the same area, but differ in *accuracy*,
2. They may be about the same area, but differ in choice of *representation*,
3. They may be about adjacent areas, and have to be *merged* into a single data set.
4. They may be about the same or adjacent areas, but referenced in different *coordinate systems*.

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

Differences in accuracy

Issues relating to positional error were outlined in Section 5.2.2, while attribute accuracy and temporal accuracy issues were discussed in Sections 5.2.3 and 5.2.4 respectively. These are clearly relevant in any combination of data sets which may themselves have varying levels of 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 readability of the map. For instance, the course of a river will only be approximated roughly on a small-scale map, and a village on its northern bank should be depicted north of the river, even if this means it has to be displaced on the map a little bit. The small scale causes an accuracy error. If we want to combine a digitized version of that map, with a digitized version of a large-scale map, we must be aware that features may not be where they seem to be. Analogous examples can be given for images at different resolutions.

Scale

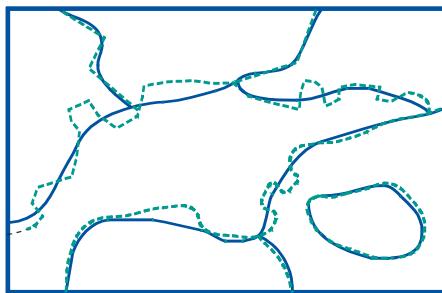


Figure 5.10: The integration of two vector data sets, which represent the same phenomenon, may lead to sliver polygons

In Figure 5.10, the polygons of two digitized maps at different scales are over-

laid. 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.

Sliver polygons

There can be good reasons for having data sets at different scales. A good example is found in mapping organizations; European organizations maintain a single source database that contains the *base data*. This database is essentially scale-less and contains all data required for even the largest scale map to be produced. For each map scale that the mapping organization produces, they derive a separate database from the foundation data. Such a derived database may be called a *cartographic* database 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.

Foundation or base data

Differences in representation

We have already talked about the various ways to represent spatial data. Sometimes data is acquired as point samples or observations, other times it is in the form of polygons with attribute data. When points need to be translated into rasters, we need to perform something known as *point data transformation*, which is discussed in Section 5.4.

Some advanced GIS applications require the possibility of representing the same geographic phenomenon in different ways. These are called *multirepresentation systems*. The production of maps at various scales is 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 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. 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.

Multi-scale and
multirepresentation systems

The links between various representations for the same object maintained by the system allows switching between them, and many fancy applications of their use seem possible. A comparison is illustrated in Figure 5.11.

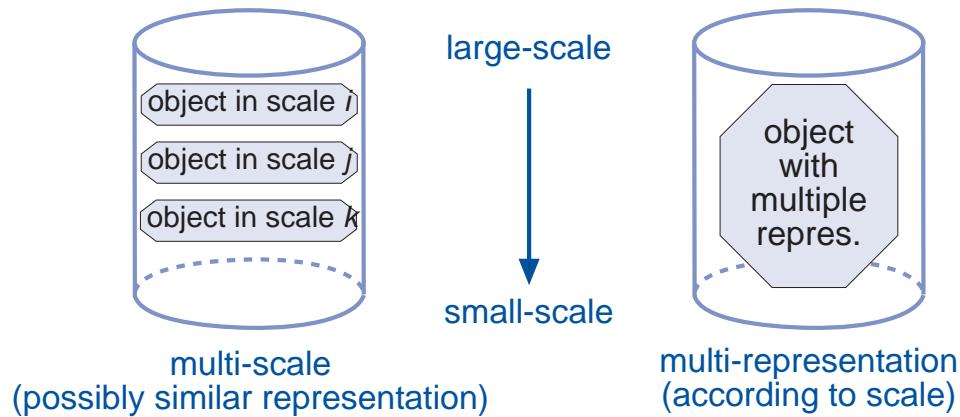


Figure 5.11: Multi-scale and multi-representation systems compared; the main difference is that multi-representation systems have a built-in ‘understanding’ that different representations belong together.

Merging data sets of adjacent areas

When individual data sets have been prepared as described above, they sometimes have to be matched into a single ‘seamless’ data set, whilst ensuring that the appearance of the integrated geometry is as homogeneous as possible. *Edge matching* is the process of joining two or more map sheets, for instance, after they have separately been digitized.

Edge matching

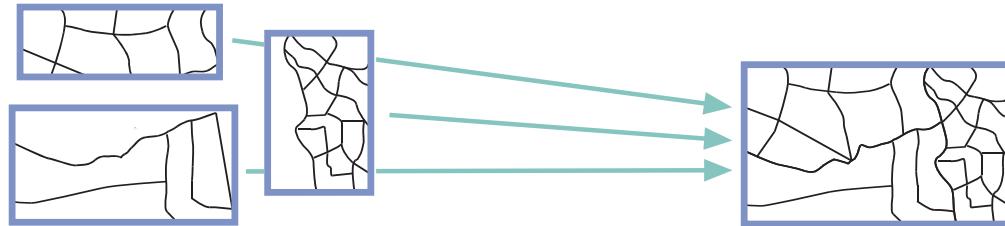


Figure 5.12: Multiple adjacent data sets, after cleaning, can be matched and merged into a single one.

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 5.12 illustrates a typical situation. Some GISs have merge or edge-matching functions to solve the problem arising from merging adjacent data. At the map sheet edges, feature representations have to be matched in order for them to be combined. Coordinates of the objects along shared borders are adjusted to match those in the neighbouring data sets. Mismatches may still occur, so a visual check, and interactive editing is likely to be required.

Differences in coordinate systems

Chapter 4 provided an introduction to coordinate systems, datums and map projections. Map projections provide means to map geographic coordinates onto a flat surface (for map production), and *vice versa*. It may be the case that data layers which are to be combined or merged in some way are referenced in different coordinate systems, or are based upon different datums. As a result, data may need coordinate transformation (Figure 4.22), or both a coordinate transformation and datum transformation (Figure 4.23). It may also be the case that data has been digitized from an existing map or data layer (Section 5.1.2). In this case, *geometric transformations* help to transform device coordinates (coordinates from digitizing tablets or screen coordinates) into world coordinates (geographic coordinates, meters, etc.).

Transformations

Other data preparation functions

A range of other data preparation functions exist that support conversion or adjustment of the acquired data to format requirements that have been defined for data storage purposes. These include:

- *Format transformation functions.* These convert between data formats of different systems or representations, e.g. reading a DXF file into a GIS. Although we will not focus on the technicalities here, the user should be warned that *conversions from one format to another may cause problems*. 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 needs it in format G (for instance because the locally preferred GIS package requires it), then usually a conversion function can be found, often within the same GIS software package. The key to successful conversion 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.
- *Graphic element editing.* Manual editing of digitized features so as to correct errors, and to prepare a clean data set for topology building.
- *Coordinate thinning.* A process that is often applied to remove redundant or excess vertices from line representations, as obtained from digitizing.

5.4 Point data transformation

This section looks at several methods of transforming point data in a GIS. We may have captured a sample of points (or acquired a dataset of such points), but wish to derive a value for the phenomenon at another location or for the whole extent of our study area.

We may want to transform our points into other representations in order to facilitate interpretation and/or integration with other data. Examples include defining homogeneous areas (polygons) from our point data, or deriving contour lines. This is generally referred to as *interpolation*, i.e. the calculation of a value from ‘surrounding’ observations. The principle of spatial autocorrelation plays a central part in the process of interpolation (see Section 2.3).

Interpolation

In order to predict the value of a point for a given (x, y) location, we could simply find the ‘nearest’ known value to the point, and assign that value. This is the simplest form of interpolation, known as *nearest-neighbour* interpolation. We might instead choose to use the distance that points are away from (x, y) to weight their importance in our calculation.

In some instances we may be dealing with a data type that limits the type of interpolation we can do (refer to page 75 for a brief background). A fundamental issue in this respect is what kind of phenomena we are considering: is it a *discrete* field—such as geological units, for instance—in which the values are of a qualitative nature and the data is categorical, or is it a *continuous* field—like elevation, temperature, or salinity—in which the values are of a quantitative nature, and represented as continuous measurements? This distinction matters,

Data type

because we are limited to nearest-neighbour interpolation for discrete data.²

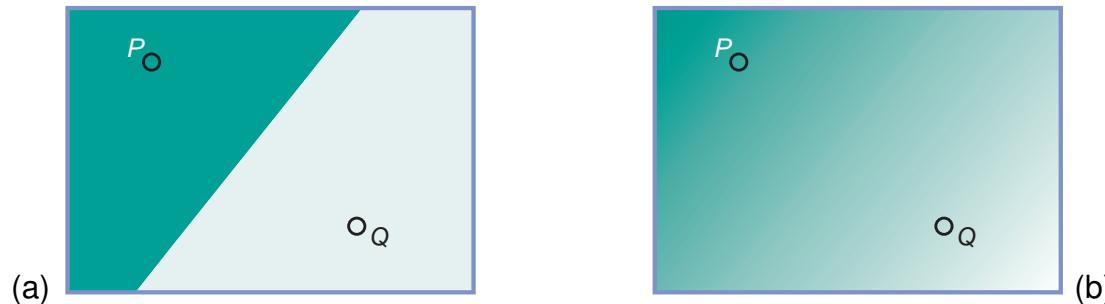


Figure 5.13: A geographic field representation obtained from two point measurements: (a) for qualitative (categorical), and (b) for quantitative (continuous) point measurements. The value measured at P is represented as dark green, that at Q as light green.

A simple example is given in Figure 5.13. Our field survey has taken only two measurements, one at P and one at Q . The values obtained in these two locations are represented by a dark and light green tint, respectively. If we are dealing with qualitative data, 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 quantitative, we can let the values of P and Q both contribute 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 point data transformation.

²Please refer to Section 2.2.3 for a background discussion of both discrete and continuous fields.

How we represent a field constructed from point measurements in the GIS also depends on the above distinction. A *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 *continuous field* can be represented as an unclassified raster, as an isoline (thus, vector) data layer, or perhaps as a TIN. Some GIS software only provide the option of generating raster output, requiring an intermediate step of raster to vector conversion. The choice of representation depends on what will be done with the data in the analysis phase.

Discrete and continuous fields

5.4.1 Interpolating discrete data

If we are dealing with discrete (nominal, categorical or ordinal) data, we are effectively restricted to using nearest-neighbour interpolation. This is the situation shown in Figure 5.13(a), though usually we would have many more points. In a nearest-neighbour interpolation, each location is assigned the value of the closest measured point. Effectively, this technique will construct ‘zones’ around the points of measurement, with each point belonging to a zone assigned the same value. Effectively, this represents an assignment of an existing value (or category) to a location.

Nearest-neighbour
interpolation

If the desired output was a polygon layer, we could construct *Thiessen polygons* around the points of measurement. The boundaries of such polygons, by definition, are the locations for which more than one point of measurement is the closest point. An illustration is provided in Figure 5.14. Thiessen polygons are further discussed on page 398. If the desired output was in the form of a raster layer, we could rasterize the Thiessen polygons. This was discussed in Section 5.3.1.

Thiessen polygons

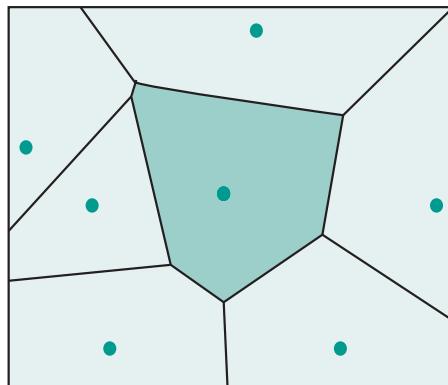


Figure 5.14: Generation of Thiessen polygons for qualitative point measurements. The measured points are indicated in dark green; the darker area indicates all locations assigned with the measurement value of the central point.

5.4.2 Interpolating continuous data

Interpolation of values from continuous measurements is significantly more complex. This is the situation of Figure 5.13(b), but again, usually with many more point measurements.

Since the data are continuous, we can make use of measured values for interpolation. There are many continuous geographic fields—elevation, temperature and ground water salinity are just a few examples. Commonly, continuous fields are represented as rasters, and we will almost by default assume that they are. Alternatives exist though, as we have seen in discussions in Chapter 2. The main alternative for continuous field representation is a polyline vector layer, in which the lines are isolines. We will also address these issues of representation below.

The aim is to use measurements to obtain a representation of the entire field using point samples. In this section we outline four techniques to do so:

1. Trend surface fitting using *regression*,
2. Triangulation,
3. Spatial moving averages using *inverse distance weighting*,
4. Kriging.

Trend surface fitting

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

The key objective in trend surface fitting is to derive a 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. Statistical techniques known as *regression techniques* can be used to determine values for these coefficients c_i that best fit with the measurements. A plane will be fitted through the measurements that makes the smallest overall error with respect to the original measurements.

Regression

In Figure 5.15, we have used the same set of point measurements, with four different approximation functions. Part (a) has been determined under the assumption that the field can be approximated by a tilted plane, in this case with a downward slope to the southeast. The values found by regression techniques were: $c_1 = -1.83934$, $c_2 = 1.61645$ and $c_3 = 70.8782$, giving us:

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

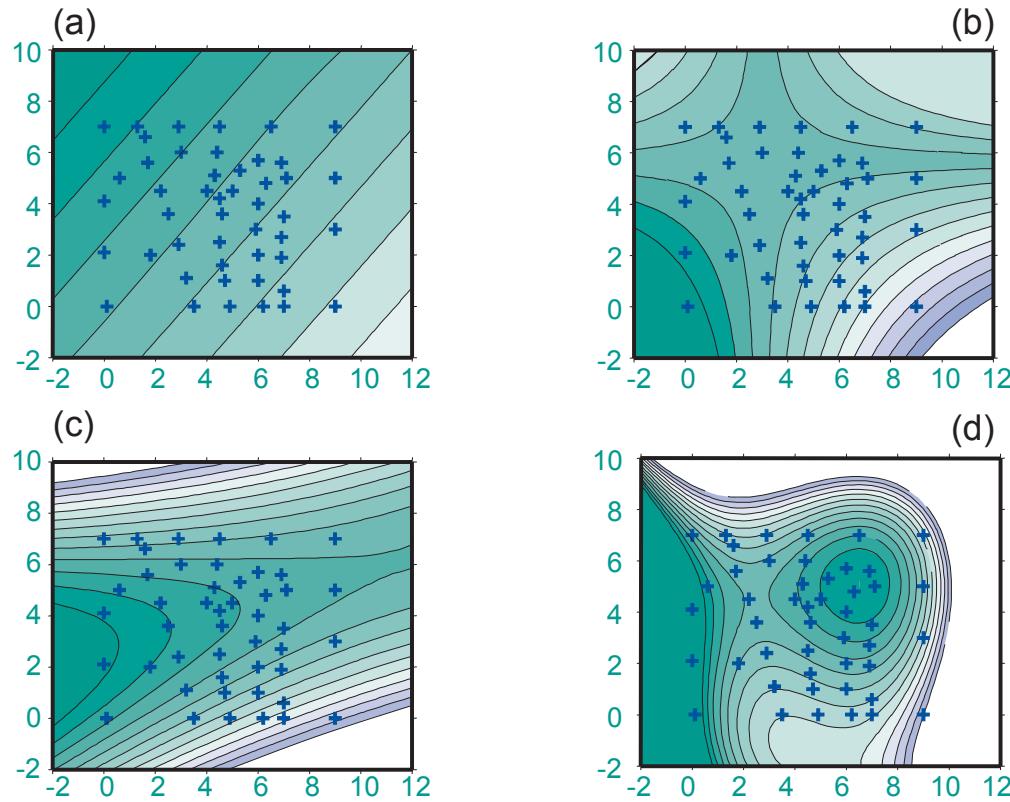


Figure 5.15: Various global trend surfaces obtained from regression techniques: (a) simple tilted plane; (b) bilinear saddle; (c) quadratic surface; (d) cubic surface. Values range from white (low), via blue, and light green to dark green (high).

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. Three such functions were the basis for the fields illustrated in Figure 5.15(b)–(d).

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

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

This 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 objective is to 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 5.15(b) and (c), respectively.

Part (d) of the figure illustrates the most complex formula of the surfaces in Figure 5.15, the *cubic surface*. It is characterized by the following formula:

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

The regression techniques applied for Figure 5.15 determined the following values for the coefficients c_i :

Fig 5.15	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, making it possible to compute an approximated value for any location in the study area.

It is possible to use trend surfaces for both *global* and *local* trends. Global trend surface fitting is based on the assumption that the entire study area can be approximated by the same mathematical surface. However in many cases, the assumption 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 quickly become computationally impossible to decipher.

Global and local trend surfaces

It should also be noted that the spatial distribution of sample measures have a significant effect on the shape of the fitting function. This is especially true for locations that are within the study area, but outside of the area within which the measurements fall. These may be subject 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 5.15. For these reasons, it is often useful to partition the study area into parts that may actually be polynomially approximated. The de-

Edge effect

cision of how to partition the study area must be taken with care, and must be guided by domain expertise. Once we have identified the parts, we may apply the trend surface fitting techniques discussed earlier, and obtain an approximation polynomial for each part.

Local trend surface fitting is not a popular technique in practical applications, because they are relatively difficult to implement, and other techniques such as moving windows are better for the representation and identification of local trends.

If we know the polynomial, it is relatively simple to generate a raster layer, given an appropriate cell resolution and an approximation function for the cell's value. In some cases it is more accurate to assign the average of the computed values for all of the cell's corner points. In order to generate a vector layer representing this data, *isolines* can be derived, for a given set of intervals. The specific techniques of generating isolines are not discussed here, however, triangulation techniques discussed below can play a role.

Generating trend surfaces

Triangulation

Another way of interpolating point measurements is by triangulation. Triangulated Irregular Networks (TINs) have already been discussed in some detail in Section 2.3.3. Essentially, this technique constructs a triangulation of the study area from the known measurement points. Preferably, the triangulation should be a *Delaunay triangulation*.³ 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, and so on. For each edge of a triangle, a geometric computation can be performed that indicates which isolines intersect it, and at what positions they do so. A list of computed locations, all at the same field value, is used by the GIS to construct the isoline. This is illustrated in Figure 5.16.

TINs and isolines

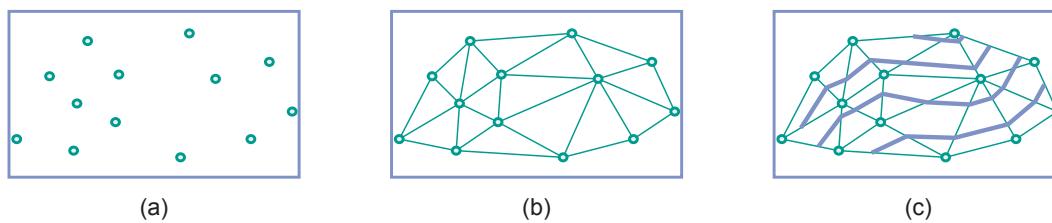


Figure 5.16: Triangulation as a means of interpolation. (a) known point measurements; (b) constructed triangulation on known points; (c) isolines constructed from the triangulation.

³For more information on this type of triangulation, see Section 6.4.1

Moving averages using inverse distance weighting (IDW)

Moving window averaging attempts to directly derive a raster dataset from a set of sample points. This is why it is sometimes also called ‘gridding’. The principle behind this technique is illustrated in Figure 5.17. The cell values for the output raster are computed one by one. To achieve this, a ‘window’ (also known as a kernel) 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. This is why moving window averaging is said to be a *local* interpolation method. 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.

Moving window averaging

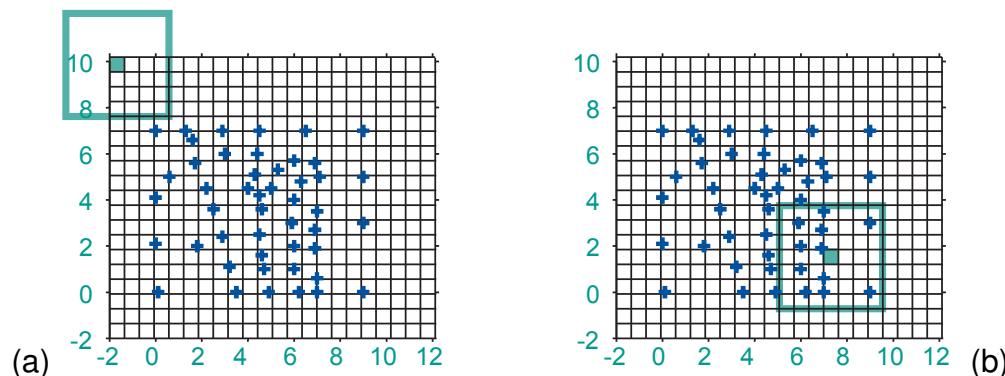


Figure 5.17: The principle of moving window averaging. In blue, the measurement points. A virtual window is moved over the raster cells one by one, and some averaging function computes a field value for the cell, using measurements within the window.

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’.

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 arithmetic mean, treating all measurements equally:

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

The principle of spatial autocorrelation suggests that measurements closer to the cell centre should have greater influence on the predicted value than those further away. In order to account for this, a distance factor can be brought into the averaging function. Functions that do this are called *inverse distance weighting functions* (IDW). This is one of the most commonly used functions in interpolating spatial data.

Weighted distance 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, but in the general case the formula is:

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

Moving window averaging has many parameters. As experimentation with any GIS package will demonstrate, picking the right parameter settings may make quite a difference for the resulting raster. We discuss some key parameters below.

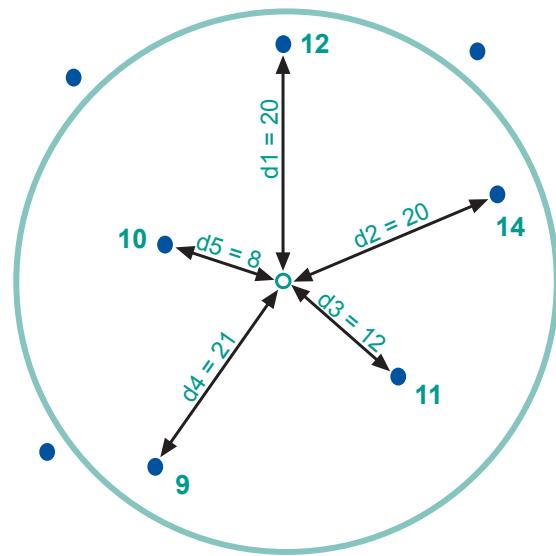


Figure 5.18: Inverse distance weighting as an averaging technique. In green, the (circular) moving window and its centre. In blue, the measurement points with their values, and distances to the centre; some are inside, some are outside of the window.

- *Raster resolution:* Too large a cell size will smooth the function too much, removing local variations; too small a 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 also possible. These can be useful 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 values, while large windows have a smoothing effect on the predicted field values.
- *Selection criteria:* Not necessarily all measurements within the window need to be used in averaging. We may choose to select use at most five, (nearest) measurements, or we may choose to only generate a field value if more than three measurements are in the window.⁴
- *Averaging function:* A final choice is which function is applied to the selected measurements within the window. It is possible to use different distance-weighting functions, each of which will influence the calculation of the resulting value.

Key IDW parameters

In many practical cases, one will have to experiment with parameter settings

⁴If 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 all directions are represented in the cell's computed value.

to obtain optimal results. When working with time series measurements (measurement sets at different points in time), one should keep the same parameter settings between time instants, as otherwise comparisons between fields computed for different moments in time will make little sense.

Kriging

Kriging was originally developed by mining geologists attempting to derive accurate estimates of mineral deposits in a given area from limited sample measurements. It is an advanced interpolation technique belonging to the field of *geostatistics*, which can deliver good results if applied properly and with enough sample points. Kriging is usually used when the variation of an attribute and/or the density of sample points is such that simple methods of interpolation may give unreliable predictions.

Kriging is based on the notion that the spatial change of a variable can be described as a function of the distance between points. It is similar to IDW interpolation, in that the surrounding values are weighted to derive a value for an unmeasured location. However, the kriging method also looks at the overall spatial arrangement of the measured points and the spatial correlation between their values, to derive values for an unmeasured location.

The first step in the kriging procedure is to compare successive pairs of point measurements to generate a *semi-variogram*. In the second step, the semi-variogram is used to calculate the weights used in interpolation. Although kriging is a powerful technique, it should not be applied without a good understanding of geostatistics, including the principle of spatial autocorrelation. For more detail on the various kriging methods, readers are referred to [11].

Semi-variogram

Discussion

The interpolation functions discussed above are available in most GISs, though each may have slightly different formulations or implementations. These are a set of the most commonly used interpolation functions, but by no means the only functions that exist.

It should be noted that there is no single *best* interpolation method, since each method has advantages and disadvantages in particular contexts. As a general guide, the following questions should be considered in selecting an appropriate method of interpolation:

- For what type of *application* will the results be used?
- What *data type* is being interpolated (e.g. categorical or continuous)?
- What is the *nature* of the surface (for example, is it a 'simple' or complex surface)?
- What is the *scale and resolution* of the data (for example, the distance between sample points)?

It is important to carry out an evaluation of the data set before interpolation takes place. In such an evaluation, one of the main goals is to establish whether there are any existing trends in the data set that may influence interpolation. Trend surfaces can be fitted to the existing data (see page 326), followed by an examination of the differences between the existing data and the resulting trend surface. It is also important to assess the spatial variability of the existing data.

Pre-interpolation checks

This can be achieved with simple moving window techniques or some other kind of linear interpolation. Finally, in order to establish the effect of the interpolation parameters on the result, different sets of interpolation parameters can be employed, and the results of these compared.

From the discussions above, an appropriate interpolation method and parameters can be determined. One way to evaluate results is to use an independent (reference) data set and calculate the difference between the value from this data set and the interpolated surface at each location. However, 'independent' datasets do not always exist. In this case, another option is to run a series of interpolations, leaving out one sample point from the original data for each run. Again, this makes it possible to compare the results from interpolation with a *known value*. If the differences ('errors') found using this method are unacceptable, either there are not enough sample points for an accurate result, or one or more of the parameters used for the interpolation is incorrect.

Evaluating interpolation
results

Summary

Digital data can be obtained directly from spatial data providers, or from pre-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. In all of these cases, data quality is a key consideration.

Data cleaning and preparation involves checking for errors, inconsistencies, and 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 have discussed various methods to address these issues in this chapter.

It is often the case that we have captured a sample of points, but wish to derive a value for the phenomenon at another location or for the whole extent of our study area. This chapter has discussed a range of point interpolation methods which can be used to achieve this. While there is no single best method, key issues to be considered in choosing the appropriate interpolation method include the application for which the data will be used, the type of data we are dealing with, the nature of the surface which the data is describing, and the scale and resolution of the data set.

Questions

1. 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.
2. 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.
3. Take another look at Figure 5.15 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 $(12, 10)$.
4. Figure 5.18 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?

Chapter 6

Spatial data analysis

The discussion up until this point has sought to prepare the reader for the ‘data analysis’ phase. So far, we have discussed the nature of spatial data, georeferencing, notions of data acquisition and preparation, and issues relating to data quality and error.

Before we move on to discuss a range of analytical operations, we should begin with some clarifications. We know from preceding discussions that the analytical capabilities of a GIS use spatial and non-spatial (attribute) data to answer questions and solve problems that are of spatial relevance. It is important to make a distinction between analysis (or analytical operations) as discussed in Section 3.3.3, and *analytical models* (often just referred to just as ‘models’). By *analysis* we mean only a subset of what is usually implied by the term: we do not specifically deal with statistical analysis (such as cluster detection, for exam-

ple). These are advanced concepts and techniques which are outside the scope of this book.

All knowledge of the world is based on models of some kind - whether they are simple abstractions, culturally-based stereotypes or complex equations that describe a physical phenomena. We have already seen in Section 1.2.1 that there are different types of *model*, and that the word itself means different things in different contexts. Section 2.1 noted that even spatial data is itself is a kind of 'model' of some part of the real world.

In this chapter we will focus on analytical functions that can form the building blocks for application models. It will hopefully become clear to the reader that these operations can be combined in various ways for increasingly complex analyses. Later in the chapter we present an overview of different types of analytical models and related concepts of which the user should be aware, as well as an examination of how various errors may degrade the results of our models or analyses.

6.1 Classification of analytical GIS capabilities

There are many ways to classify the analytical functions of a GIS. The classification used for this chapter, is essentially the one put forward by Aronoff [3]. It makes the following distinctions, which are addressed in subsequent sections of the chapter:

1. **Classification, retrieval, and measurement functions.** All functions in this category are performed on a single (vector or raster) data layer, often using the associated attribute data.
 - Classification allows the assignment of features to a class on the basis of attribute values or attribute ranges (definition of data patterns). On the basis of reflectance characteristics found in a raster, pixels may be classified as representing different crops, such as potato and maize.
 - Retrieval functions allow the selective search of data. We might thus retrieve all agricultural fields where potato is grown.
 - Generalization is a function that joins different classes of objects with common characteristics to a higher level (generalized) class.¹ For ex-

¹The 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.

ample, we might generalize fields where potato or maize, and possibly other crops, are grown as ‘food produce fields’.

- Measurement functions allow the calculation of distances, lengths, or areas.

More detail can be found in Section 6.2.

2. **Overlay functions.** These belong to the most frequently used functions in a GIS application. They allow the combination of two (or more) spatial data layers comparing them position by position, and treating areas of overlap—and of non-overlap—in distinct ways. Many GISs support overlays through an algebraic language, expressing an overlay function as a formula in which the data layers are the arguments. In this way, we can find

- The potato fields on clay soils (select the ‘potato’ cover in the crop data layer and the ‘clay’ cover in the soil data layer and perform an *intersection* of the two areas found),
- The fields where potato or maize is the crop (select both areas of ‘potato’ and ‘maize’ cover in the crop data layer and take their *union*),
- The potato fields not on clay soils (perform a *difference* operator of areas with ‘potato’ cover with the areas having clay soil),
- The fields that do not have potato as crop (take the *complement* of the potato areas).

These are discussed further in Section 6.3.

3. **Neighbourhood functions.** Whereas overlays combine features at the same location, neighbourhood functions evaluate the characteristics of an area *surrounding* a feature's location. A neighbourhood function 'scans' the neighbourhood of the given feature(s), and performs a computation on it.

- *Search functions* allow the retrieval of features that fall within a given *search window*. This window may be a rectangle, circle, or polygon.
- *Buffer zone generation* (or buffering) is one of the best known neighbourhood functions. It determines a spatial envelope (*buffer*) around (a) given feature(s). The created buffer may have a fixed width, or a variable width that depends on characteristics of the area.
- *Interpolation functions* predict unknown values using the known values at nearby locations. This typically occurs for continuous fields, like elevation, when the data actually stored does not provide the direct answer for the location(s) of interest. Interpolation of continuous data was discussed in Section 5.4.2.
- *Topographic functions* determine characteristics of an area by looking at the immediate neighbourhood as well. Typical examples are slope computations on digital terrain models (i.e. continuous spatial fields). The *slope* in a location is defined as the plane tangent to the topography in that location. Various computations can be performed, such as:
 - determination of *slope angle*,
 - determination of *slope aspect*,
 - determination of *slope length*,

- determination of *contour lines*. These are lines that connect points with the same value (for elevation, depth, temperature, barometric pressure, water salinity etc).

We discuss these topics more fully in Section 6.4.

4. **Connectivity functions.** These functions work on the basis of networks, including road networks, water courses in coastal zones, and communication lines in mobile telephony. These networks represent spatial linkages between features. Main functions of this type include:

- *Contiguity functions* evaluate a characteristic of a set of connected spatial units. One can think of the search for a contiguous area of forest of certain size and shape in a satellite image.
- *Network analytic functions* are used to compute over connected line features that make up a network. The network may consist of roads, public transport routes, high voltage lines or other forms of transportation infrastructure. Analysis of such networks may entail *shortest path computations* (in terms of distance or travel time) between two points in a network for routing purposes. Other forms are to find all points reachable within a given distance or duration from a start point for allocation purposes, or determination of the capacity of the network for transportation between an indicated source location and sink location.
- *Visibility functions* also fit in this list as they are used to compute the points visible from a given location (viewshed modelling or viewshed mapping) using a digital terrain model.

Details are discussed in Section 6.5.

6.2 Retrieval, classification and measurement

6.2.1 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. In general, measurements on vector data are more advanced, thus, also more complex, than those on raster data. We discuss each group.

Measurement types

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.

The attentive reader will have noted that all of the above ‘measurements’ do not actually require computation, but only *retrieval* of stored data.

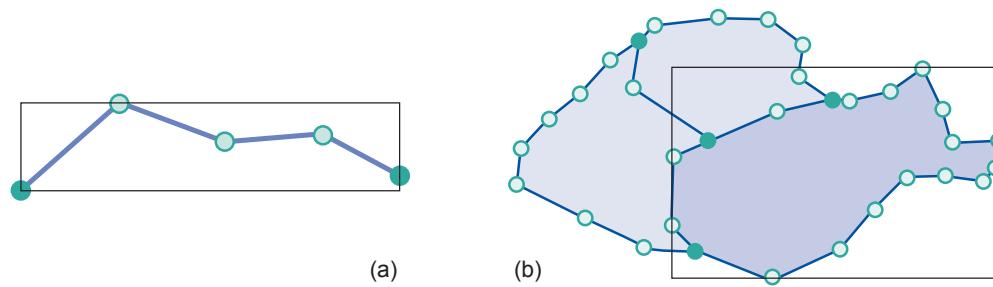
Measuring distance between two features is another important function. If both features are points, say p and q , the computation in a Cartesian spatial reference system are given by the well-known Pythagorean distance function:

$$\text{dist}(p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}.$$

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. It is not possible to store all distance values for all possible combinations of two features in any reasonably sized spatial database. As a result, the system must compute 'on the fly' whenever a distance computation request is made.

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 6.1. 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 a complicated 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.

Minimal bounding box

**Figure 6.1:** The minimal bounding box of (a) a polyline, and (b) a polygon

For practical purposes, it is important to be aware of the measurement unit that applies to the spatial data layer that one is working on. This is determined by the spatial reference system that has been defined for it during data preparation.

A common use of area size measurements is when one wants to sum up the area sizes of all polygons belonging to some class. This class could be crop type: What is the size of the area covered by potatoes? If our crop classification is in a stored data layer, the computation would include (a) selecting the potato areas, and (b) summing up their (stored) area sizes. Clearly, little *geometric* computation is required in the case of stored features. This is not the case when we are interactively defining our vector features in GIS use, and we want measurements to be performed on these interactively defined features. Then, the GIS will have to perform complicated geometric computations.

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 by 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 sum of distances between consecutive cells. This computation is prone to error, as already discovered in Chapter 2 (Question 11).

6.2.2 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.

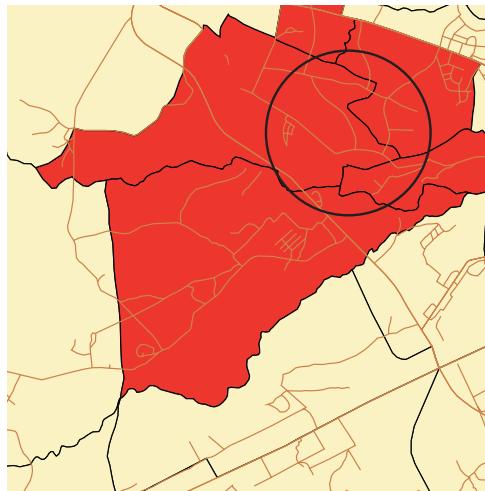
Interactive spatial selection

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 2.15) with the selection objects. These become the *selected objects*.

Selection objects

As we have seen in Section 3.5, spatial data stored in a geodatabase is associated with its attribute data through a key/foreign key link. Selections of features lead 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 6.2, the selection object is a circle and the selected objects are the red polygons; they overlap with the selection object.

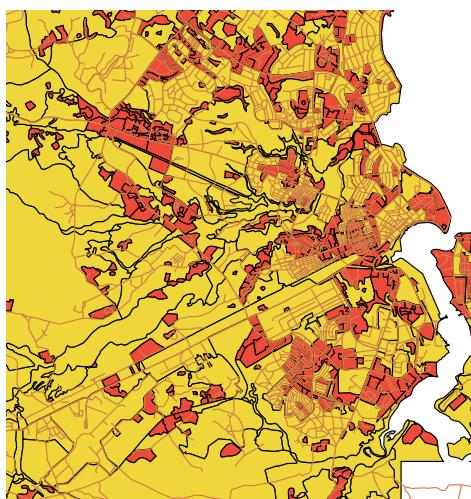


Area	Perimeter	Ward_id	Ward_name	District	Pop88	Pop92
65420380.0000	41654.940000	1	KUNDUCHI	Kinondoni	22106	27212.00
24813620.0000	30755.620000	2	KAWE	Kinondoni	32854	40443.00
18698500.0000	26403.580000	3	MSASANI	Kinondoni	51225	63058.00
81845610.0000	49645.160000	4	UBUNGO	Kinondoni	47281	58203.00
446954610.0000	13480.130000	5	MANZESE	Kinondoni	59467	73204.00
4999599.0000	10356.850000	6	TANDALE	Kinondoni	58357	71837.00
4102218.0000	8951.096000	7	MWANANYAMALA	Kinondoni	72956	88809.00
3749840.0000	9447.422000	8	KINONDONI	Kinondoni	42301	52073.00
2087509.0000	7502.250000	9	UPANGA WEST	Ilala	9852	11428.00
2288613.0000	9028.789000	10	KIVUKONI	Ilala	5391	6254.00
1400024.0000	6883.288000	11	NDUGUMBI	Kinondoni	32548	40067.00
888966.900000	4589.110000	12	MAGOMENI	Kinondoni	16938	20851.00
1448370.0000	5651.958000	13	UPANGA EAST	Ilala	11019	12782.00
6214379.0000	14652.080000	14	MABIBO	Kinondoni	43381	53402.00
2499922.0000	7121.359000	15	MAKURUMILA	Kinondoni	54141	66648.00
1262028.0000	4885.793000	16	MIZIMUNI	Kinondoni	23989	29530.00
35362240.0000	28976.090000	17	KINYEREZI	Ilala	3044	3531.00
1010813.0000	5393.771000	18	JANGIWANI	Ilala	15297	17745.00
475745.500000	3043.069000	19	KISUTU	Ilala	8399	9743.00
1754043.0000	7743.187000	20	KIGOGO	Kinondoni	21267	26160.00
29964950.0000	36964.000000	21	KIGAMBONI	Temeke	23203	27658.00
1291479.0000	5187.690000	22	MICHIKICHINI	Ilala	14852	17228.00
720322.100000	4342.732000	23	MCHAFUKOGE	Ilala	8439	9789.00
6999131.0000	16321.530000	24	TABATA	Ilala	18454	21407.00
483620.700000	3304.072000	25	KARIAKOO	Ilala	12506	14507.00
3564653.0000	9586.751000	26	BUGURUNI	Ilala	48286	56012.00
2639375.0000	6970.189000	27	ILALA	Ilala	35372	41032.00
912452.800000	4021.937000	28	GEREZANI	Ilala	7490	8688.00
6735135.0000	13579.590000	29	KURASINI	Temeke	26737	31871.00

Figure 6.2: 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 in Dar es Salaam, Tanzania. Data source: Dept. of Urban & Regional Planning and Geo-information Management, ITC.

Spatial selection by attribute conditions

It is also possible to select features by using selection conditions on feature attributes. These conditions are formulated in SQL if the attribute data reside in a geodatabase. This type of selection answers questions like “where are the features with ...?”

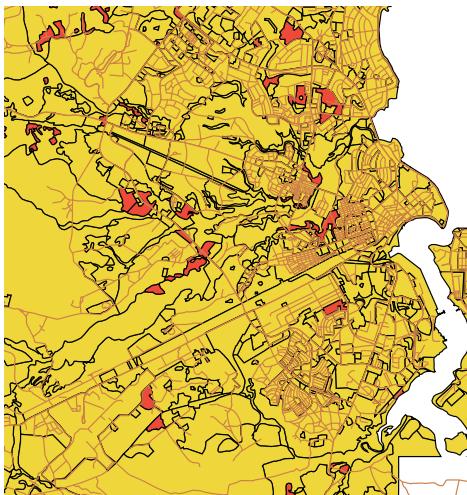


Area	IDs	LandUse
174308.70	2	30
2066475.00	3	70
214582.50	4	80
29313.86	5	80
73328.08	6	80
53303.30	7	80
614530.10	8	20
1637161.00	9	80
156357.40	10	70
59202.20	11	20
83289.59	12	80
225642.20	13	20
28377.33	14	40
228930.30	15	30
986242.30	16	70

Figure 6.3: Spatial selection using the attribute condition $\text{Area} < 400000$ on land use areas in Dar es Salaam. Spatial features on left, associated attribute data (in part) on right. Data source: Dept. of Urban & Regional Planning and Geo-information Management, ITC.

Figure 6.3 shows an example of selection by attribute condition. The query expression is $\text{Area} < 400000$, 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 this 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 se-

lected features of Figure 6.3 are subjected to a further condition, $LandUse = 80$. The result is illustrated in Figure 6.4. Such combinations of conditions are fairly common in practice, so we devote a small paragraph on the theory of combining conditions.



Area	IDs	LandUse
174308.70	2	30
2066475.00	3	70
214582.50	4	80
29313.86	5	80
73328.08	6	80
53303.30	7	80
614530.10	8	20
1637161.00	9	80
156357.40	10	70
59202.20	11	20
83289.59	12	80
225642.20	13	20
28377.33	14	40
228930.30	15	30
986242.30	16	70

Figure 6.4: Further spatial selection from the already selected features of Figure 6.3 using the additional condition $LandUse = 80$ on land use areas. Observe that fewer features are now selected. Data source: Dept. of Urban & Regional Planning and Geo-information 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 $\text{Area} < 400000$, and $\text{LandUse} = 80$. Atomic conditions use a predicate symbol, such as $<$ (less than) or $=$ (equals).

Other possibilities are \leq (less than or equal), $>$ (greater than), \geq (greater than or equal) and \neq (does not equal). Any of these symbols is combined with an expression on the left and one on the right. For instance, $\text{LandUse} \neq 80$ 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 $0.15 \times \text{Area}$, which would compute 15% of the area size.

Atomic and composite conditions

Atomic conditions can be combined into *composite conditions* using *logical connectives*. The most important ones are *AND*, *OR*, *NOT* and the bracket pair (\dots) . If we write a composite condition like

$\text{Area} < 400000 \text{ AND } \text{LandUse} = 80,$

we can use it to select areas for which *both* atomic conditions hold true. This is the meaning of the *AND* connective. If we had written

Logical connectives

$\text{Area} < 400000 \text{ OR } \text{LandUse} = 80$

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 2.3.4. These relationships can be useful to select features as well. The steps carried out are:

1. To select one or more features as the selection objects, and
2. To apply a chosen spatial relationship function to determine the selected features that have that relationship with the selection objects.

Selecting features that are inside selection objects This type of query uses the *containment relationship* between spatial objects. Obviously, polygons can contain polygons, lines or points, and lines can contain lines or points, but no other containment relationships are possible.

Point-in-polygon query

Figure 6.5 illustrates a containment query. Here, we are interested in finding the location of 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 = “Ilala”*. Then, these selected areas were used as selection objects to determine which medical clinics (as point objects) were within them.

Selecting features that intersect The intersect operator identifies features that are not disjoint in the sense of Figure 2.15, but now extended to include points and lines. Figure 6.6 provides an example of spatial selection using the intersect relationship between lines and polygons. We selected all roads intersecting Ilala District.

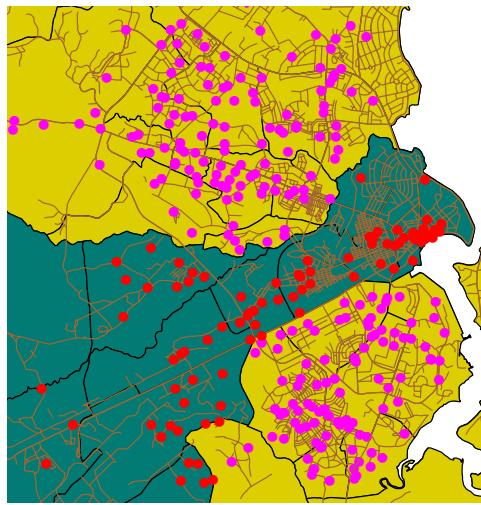


Figure 6.5: Spatial selection using containment. In dark green, all wards within Ilala District as the selection objects. In red, all medical clinics located inside these areas, and thus inside the district. Data source: Dept. of Urban & Regional Planning and Geo-information Management, ITC.

Selecting features adjacent to selection objects Adjacency is the *meet* relationship of Section 2.3.4. It expresses that features share boundaries, and therefore it applies only to line and polygon features. Figure 6.7 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, e.g.:

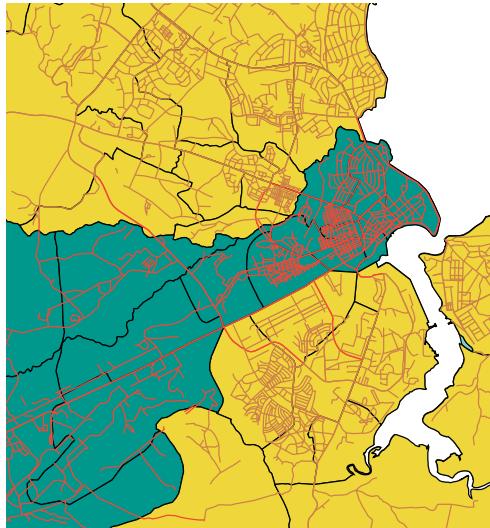


Figure 6.6: Spatial selection using intersection. The wards of Ilala District function as the selection objects (in dark green), and all roads (partially) in the district are selected (in red). Data source: Dept. of Urban & Regional Planning and Geo-information Management, ITC.

- 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 6.8 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.

In situations in which we know the distance criteria to use—for selections within, at or beyond that distance value—the GIS has many (straightforward) computations to perform. Things become more complicated if 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.

Complex proximity formulations

Afterthought on selecting features So far we have discussed a number of different techniques for selecting features. We have also seen that selection conditions on attribute values can be combined using logical connectives like *AND*, *OR* and *NOT*. A fact is that the other techniques of selecting features can usu-

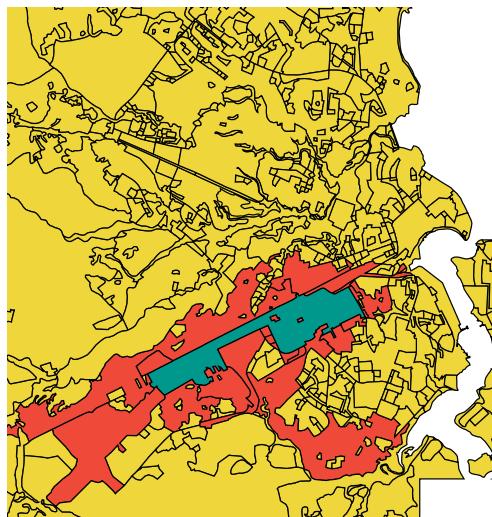


Figure 6.7: Spatial selection using adjacency. Our selection object is an industrial area near down town Dar es Salaam, Tanzania; our adjacency selection finds all adjacent land use areas. Data source: Dept. of Urban & Regional Planning and Geo-information Management, ITC.

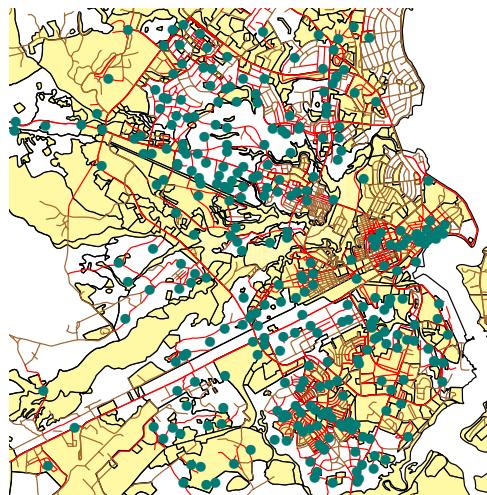


Figure 6.8: Spatial selection using the distance function. With all clinics being our selection objects, we searched for roads that pass by within 200 metres. Observe that this also selects road segments that are far away from any clinic, simply because they belong to a road of which a segment is nearby. Data source: Dept. of Urban & Regional Planning and Geo-information Management, ITC.

ally also be *combined*. Any set of selected features can be used as the input for a subsequent selection procedure. This means, for instance, that we can select all medical clinics first, then identify 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. In this way, we are combining various techniques of selection.

Combining selection conditions

6.2.3 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 6.9 in two ways.

Classification parameter

The input data set may have itself been the result of a classification, and in such a case we call it 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.

Reclassification

In classification of vector data, there are two possible results. In the first, the

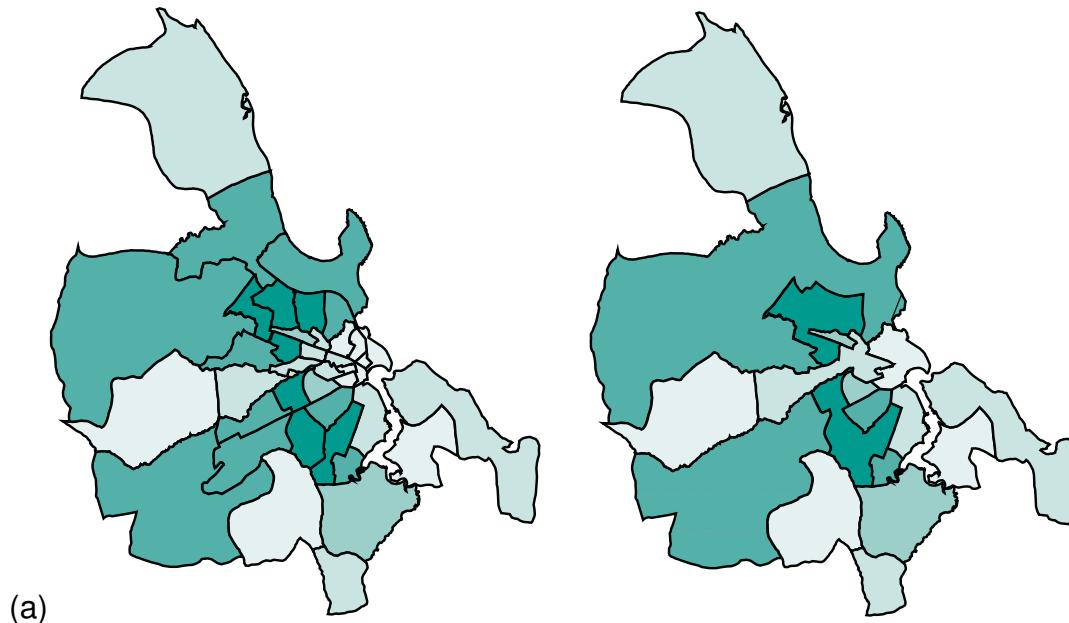


Figure 6.9: Two classifications of average annual household income per ward in Dar es Salaam, Tanzania. Higher income areas in darker greens. Five categories were identified. (a) with original polygons left intact; (b) with original polygons merged when in same category. The data used for this illustration are *not* factual.

input features may become the output features in a new data layer, with an additional category assigned. In other words, nothing changes with respect to the spatial extents of the original features. Figure 6.9(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 post-processing functions are called *spatial merging*, *aggregation* or *dissolving*. An illustration of this second type is found in Figure 6.9(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

Aggregation and merging

only for lines and polygons. Below, we discuss two kinds of classification: user-controlled and automatic.

<i>Household income range</i>	<i>New category value</i>
391–2474	1
2475–6030	2
6031–8164	3
8165–11587	4
11588–21036	5

Table 6.1: Classification table used in Figure 6.9.

User-controlled classification

In *user-controlled classification*, a user selects the attribute(s) that will be used as the classification parameter(s) and defines the classification method. The latter involves declaring the number of classes as well as the correspondence between the old attribute values and the new classes. This is usually done via a classification table. The classification table used for Figure 6.9 is displayed in Table 6.1. 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 2.2.3.

Another case exists when the classification parameter is nominal or at least discrete. Such an example is given in Figure 6.10.

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 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.

Null value

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

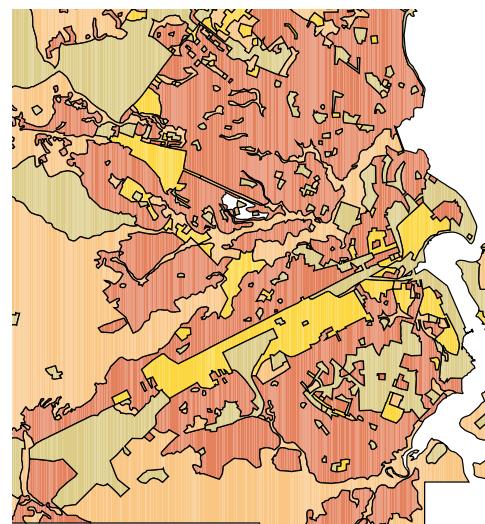


Figure 6.10: An example of a classification on a discrete parameter, namely land use unit in the city of Dar es Salaam, Tanzania. Colour scheme: Residential (brown), Commercial (yellow), Public (Olive), Non built-up (orange). Data source: Dept. of Urban & Regional Planning and Geo-information Management, ITC.

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 main techniques of determining break points are in use.

1. *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.
2. *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 6.11.

When to use which? Which of these techniques should be applied to a given dataset depends upon the purpose of the analysis (what the user is trying to

achieve) as well as the characteristics of the data itself. The reader is encouraged to experiment with their data and compare the results given by each method. Other (and possibly better) techniques exist.

While these two types of classification can be used in spatial analysis, they are also frequently used to develop *visualizations* of the same phenomena. In terms of analytical operations we refer to some kind of calculation or function which will use these categories. In terms of visualization, we refer to the graphical representation of the data using these classifications. Just as either technique yields different results in numeric terms, it will do the same in visual terms. Please refer to Chapter 7 for more discussion on issues relating to mapping and visualization.

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

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
1,2	1	9
3,4	2	8
5,6	3	3
7,8	4	3
9,10	5	2

original value	new value	# cells
1	1	6
2,3	2	5
4	3	6
5,6	4	3
8,9,10	5	5

Figure 6.11: 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.

6.3 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 from them. The binary operators that we discuss are known as *spatial overlay operators*. We will firstly discuss vector overlay operators, and then focus on the raster case.

Standard overlay operators take two input data layers, and assume they are geo-referenced in the same system, and overlap in study area. If either of these requirements 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 result for each location in the output data layer. The specific result to produce is determined by the user. It might involve a calculation, or some other logical function to be applied to every area or location.

Overlay requirements

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 applies, but the underlying computations rely on determining the spatial intersections of features from each input layer.

6.3.1 Vector overlay operators

In the vector domain, overlay is computationally more demanding than in the raster domain. Here we will only discuss overlays from polygon data layers, but we note that most of the ideas also apply to overlay operations with point or line data layers.

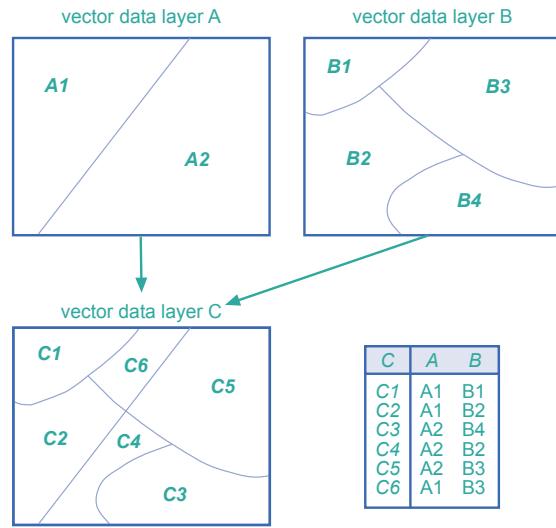


Figure 6.12: The polygon intersect (overlay) operator. Two polygon layers *A* and *B* produce a new polygon layer (with associated attribute table) that contains all intersections of polygons from *A* and *B*. Figure after [8].

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 6.12. 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

Spatial join

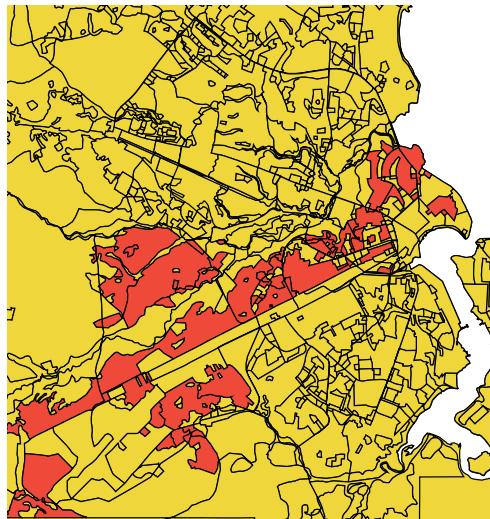


Figure 6.13: The residential areas of Ilala District, obtained from polygon intersection. Input for the polygon intersection operator were (a) a polygon layer with all Ilala wards, (b) a polygon layer with the residential areas, as classified in Figure 6.10. Data source: Dept. of Urban & Regional Planning and Geo-information Management, ITC.

attribute table only contains one tuple for each intersection polygon found, and this explains why we call this operator a *spatial join*.

A more practical example is provided in Figure 6.13, which was produced by polygon intersection of the ward polygons with land use polygons classified as in Figure 6.10. This has allowed us to select the residential areas in Ilala District.

Two more polygon overlay operators are illustrated in Figure 6.14. 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 (selected) polygons in a second input layer. Besides this generalized outer boundary, no other polygon boundaries from the second layer play a role in the result.

Polygon clipping

A second overlay operator is *polygon overwrite*. The result of this binary operator

is defined is a polygon layer with the polygons of the first layer, except where polygons existed in the second layer, as these take priority. The principle is illustrated in the lower half of Figure 6.14. 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 an operator argument.

The 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.

Polygon intersection

Vector overlays are usually also defined for point or line data layers. Their definition parallels the definitions of operators discussed above. Different GISs use different names for these operators, and one is advised to carefully check the documentation before applying any of these operators.

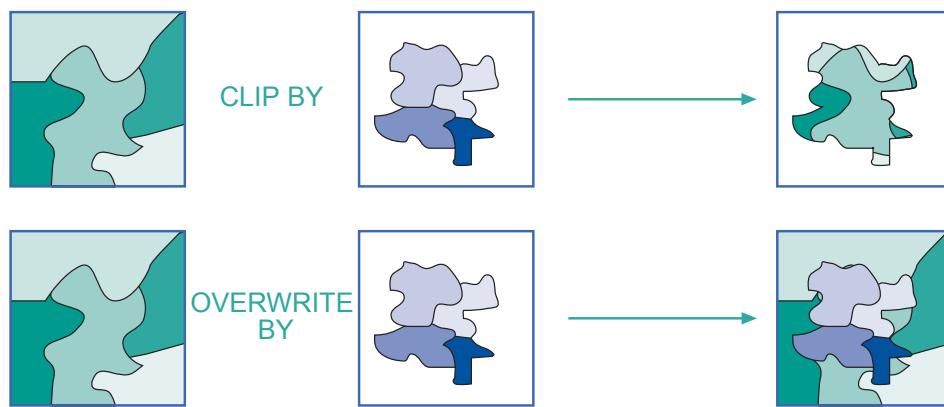


Figure 6.14: Two more polygon overlay operators:
(a) polygon clip overlay clips down the left hand polygon layer to the generalized spatial extent of the right hand polygon layer;
(b) polygon overwrite overlay overwrites the left hand polygon layer with the polygons of the right hand layer.

6.3.2 Raster overlay operators

Vector overlay operators are useful, but geometrically complicated, and this sometimes results in poor operator performance. Raster overlays do not suffer from this disadvantage, as most of them perform their computations cell by cell, and thus they are fast.

GISs that support raster processing—as most do—usually have a language to express operations on rasters. These languages are generally referred to as *map algebra* [54], or sometimes raster calculus. They allow a GIS to compute new rasters from existing ones, using a range of functions and operators. Unfortunately, not all implementations of map algebra offer the same functionality. The discussion below is to a large extent based on general terminology, and attempts to illustrate the key operations using a logical, structured language. Again, the syntax often differs for different GIS software packages.

Map algebra

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 := *Map_algebra_expression*.

The expression on the right is evaluated by the GIS, and the raster in which it results is then stored under the name on the left. The expression may contain references to existing rasters, operators and functions; the format is made clear 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

Operands

continuing until the last pixel in the last row. There is a wide range of operators and functions that can be used in map algebra, which we discuss below.

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, $10 \text{ MOD } 3$ will return 1 as $10 - 3 \times 3 = 1$. Similarly, $10 \text{ DIV } 3$ 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.

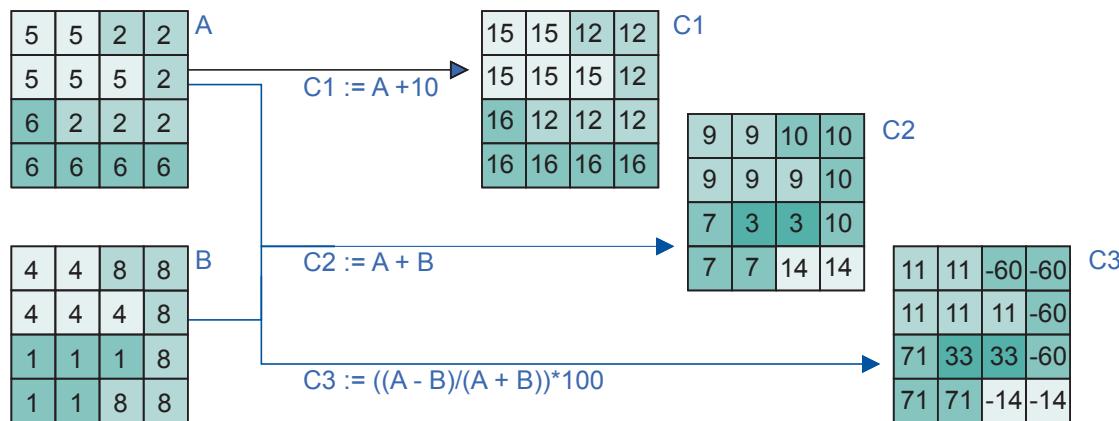


Figure 6.15: Examples of arithmetic map algebra expressions

Some simple map algebra assignments are illustrated in Figure 6.15. The assign-

ment:

$$C1 := A + 10$$

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

Map algebra also allows the comparison of rasters cell by cell. To this end, we may use the standard comparison operators ($<$, \leq , $=$, \geq , $>$ and \neq) that we introduced before.

A simple raster comparison assignment is:

$$C := A \neq B.$$

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 most implementations of map algebra. We have already seen the connectives of *AND*, *OR* and *NOT* in Section 6.2.2. Another connective that is commonly offered in map algebra is *exclusive OR (XOR)*. The expression $a \text{ XOR } b$ is true only if either a or b is true, but not both. Examples of the use of these comparison operators and connectives are provided in Figure 6.16 and Figure 6.17. The latter figure provides various raster computations in search of forests at specific elevations. In the figure, raster $D1$ indicates forest below 500 m, $D2$ indicates areas below 500 m that are forests, raster $D3$ areas that are either forest or below 500 m (but not at the same time), and raster $D4$ indicates forests above 500 m.

Comparison operators and connectives

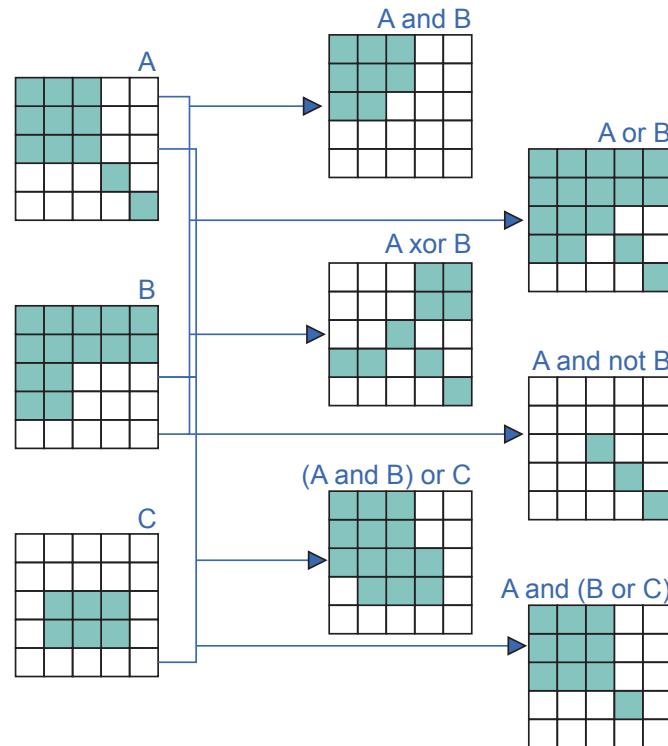


Figure 6.16: Examples of logical expressions in map algebra. Green cells represent true values, white cells represent false values.

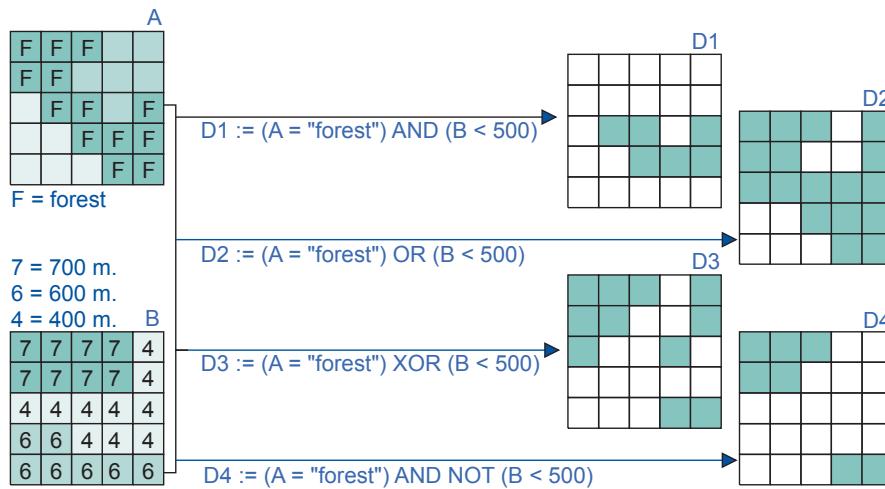


Figure 6.17: Examples of complex logical expressions in map algebra. A is a classified raster for land use, and B holds elevation values.

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 us to test whether a condition is fulfilled. The general format is:²

$$\text{Output_raster} := \text{CON}(\text{condition}, \text{then_expression}, \text{else_expression}).$$

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 $\text{CON}(A = \text{"forest"}, 10, 0)$ will evaluate to *10* for each cell in the output raster where the same cell in *A* is classified as forest. In each cell where this is not true, the *else_expression* is evaluated, resulting in *0*. Another example is provided in Figure 6.18, showing that values for the *then_expression* and the *else_expression* can be some integer (possibly derived from another calculation) or values derived from other rasters. In this example, the output raster *C1* is assigned the values of input raster *B* wherever the cells of input raster *A* contain forest. The cells in output raster *C2* are assigned *10* wherever the elevation (*B*) is equal to 7 and the groundcover (*A*) is forest.

² We have already noted that specific software packages may differ in the specifics of the syntax that make up an expression. This extends to the actual commands—some packages using “IFF” instead of “CON”.

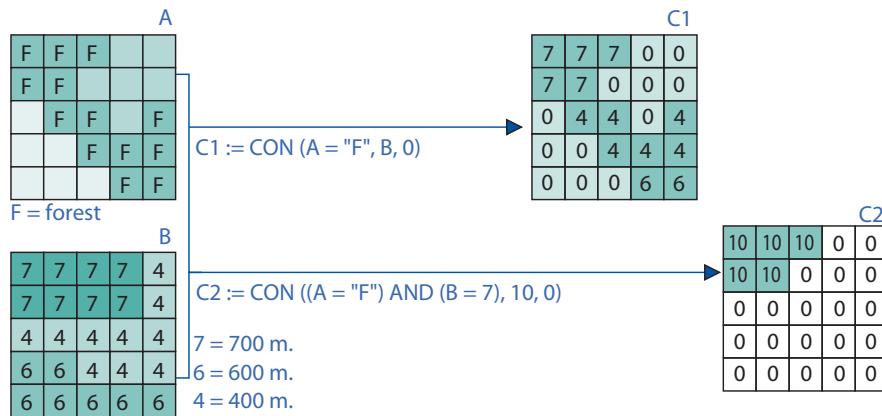


Figure 6.18: Examples of conditional expressions in map algebra. Here A is a classified raster holding land use data, and B is an elevation value raster.

6.3.3 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 6.19 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.

Domain expertise

We could produce the output raster of Figure 6.19 with a map algebra expression such as:

$$\text{Suitability} := \text{CON}((\text{Landuse} = \text{"Forest"} \text{ AND } \text{Geology} = \text{"Alluvial"}) \text{ OR} \\ (\text{Landuse} = \text{"Grass"} \text{ AND } \text{Geology} = \text{"Shale"}), \\ \text{"Suitable", "Unsuitable"})$$

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 *CON* expression is not an easy task.

To this end, some GISs accommodate setting up a separate decision table that will guide the raster overlay process. This extra table carries domain expertise,

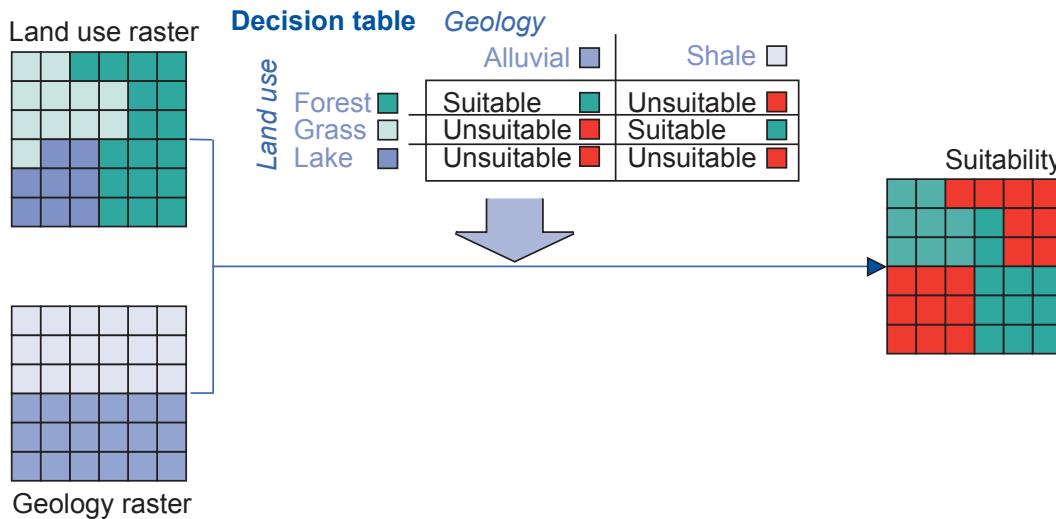


Figure 6.19: The use of a decision table in raster overlay. The overlay is computed in a suitability study, in which land use and geology are important factors. The meaning of values in both input rasters, as well as the output raster can be understood from the decision table.

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 6.19. The GIS will have supporting functions to generate the additional table from the input rasters, and to enter appropriate values in the table.

6.4 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 map algebra, 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 define 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 might be a medical clinic. Its neighbourhood could 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.

The alert reader will note the increasingly complex definitions of ‘neighbourhood’ used here. This is to illustrate that different ways of measuring neighbourhoods exist, and some are better (or more representative of *real* neighbourhoods) than others, depending on the purpose of the analysis.

Then, in the third step we indicate what it is we want to discover about the phenomena that exist or occur in the neighbourhood. This might simply be its spatial extent, but it might also be statistical information like:

- The total population of the area,
- Average household income, or
- The distribution of 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 in this case, 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.

Determining neighbourhood extent To select target locations, one can use the selection techniques that we discussed in Section 6.2.2. 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 6.4.1. Geometric distance does not take into account direction and certain phenomena can only be studied by doing so. For example, pollution spread by rivers, ground water flow, or prevailing weather systems.

The more advanced techniques for computation of flow and diffusion are discussed in Section 6.4.2. Diffusion 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 diffusion. In flow 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. Both flow and diffusion computations take local characteristics into account, and are therefore more easily performed on raster data.

Proximity function

Complex neighbourhoods

6.4.1 Proximity computations

In proximity computations, we use geometric distance to define the neighbourhood of one or more target locations. The most common and useful technique is *buffer zone generation*. Another technique based on geometric distance that 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 6.20(a), a number of main and minor roads were selected as targets, and a 75 m (resp., 25 m) buffer was computed from them. In some case studies, zonated buffers must be determined, for instance in assessments of traffic noise effects. Most GISs support this type of zonated buffer computation. An illustration is provided in Figure 6.20(b).

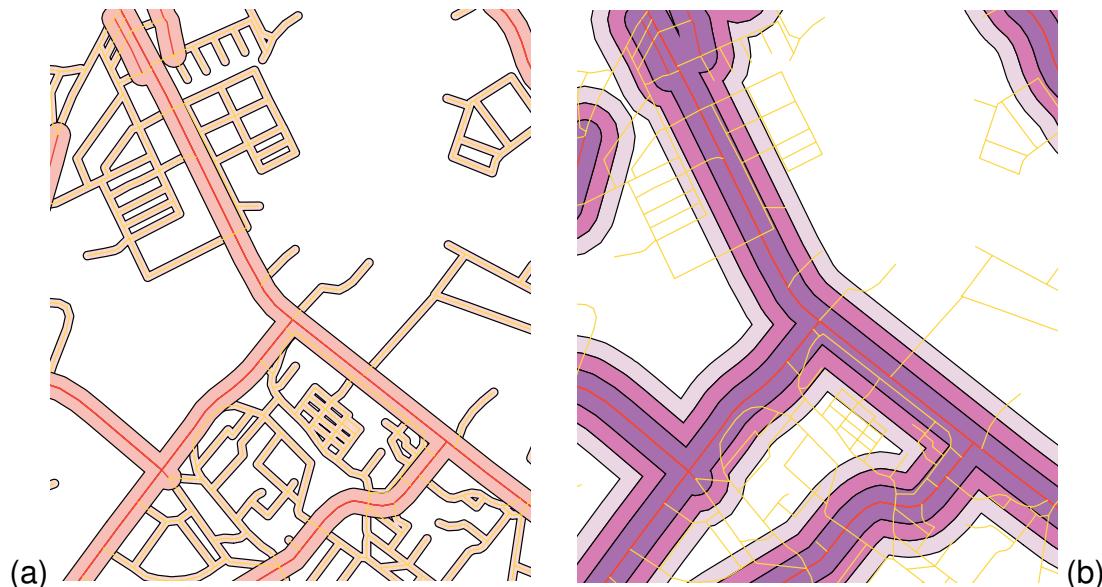


Figure 6.20: Buffer zone generation: (a) around main and minor roads. Different distances were applied: 25 metres for minor roads, 75 metres for main roads. (b) Zonated buffer zones around main roads. Three different zones were obtained: at 100 metres from main road, at 200, and at 300 metres.

In vector-based buffer generation, the buffers themselves become polygon features, usually in a separate data layer, that can be used in further spatial analysis.

Buffer generation on rasters is a fairly simple function. The target location or locations are always represented by a selection of the raster's cells, and geometric distance is defined, using cell resolution as the unit. The distance function applied is the Pythagorean distance between the cell centres. The distance from a non-target cell to the target is the minimal distance one can find between that non-target cell and any target cell.

Thiessen polygon generation

Thiessen polygon partitions make use of geometric distance for determining neighbourhoods. This 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. We have already seen the use of Thiessen polygons in the context of *interpolation* of point data, as discussed in Section 5.4.1. 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 2.3.3 on TINs).

Figure 6.21 repeats the Delaunay triangulation of Figure 2.9(b). The Thiessen polygon partition constructed from it is on the right. 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.

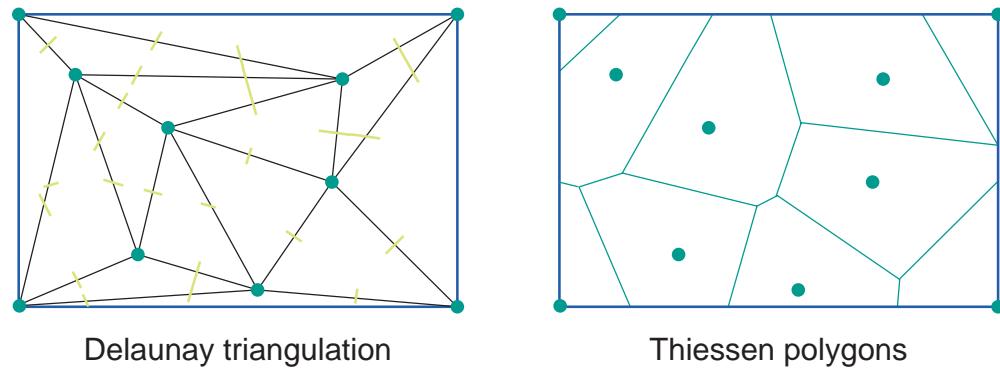


Figure 6.21: Thiessen polygon construction (right) from a Delaunay triangulation (left): perpendiculars of the triangles form the boundaries of the polygons.

6.4.2 Computation of diffusion

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, referred to as *diffusion*. 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 GISs provide support for this type of computation, and we discuss some of its principles here, in the context of raster data.

Diffusion and spread

Diffusion 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. The computation also involves a *local resistance raster*, which for each cell provides a value that indicates how difficult it is for the ‘source - material’ to pass 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. This process is illustrated in Figure 6.22.

Resistance

While computing total resistances, the GIS takes proper care of the path lengths. Obviously, the diffusion from a cell c_{src} to its neighbour cell to the east c_e is shorter than to the cell that is its northeast neighbour c_{ne} . The distance ratio between these two cases is $1 : \sqrt{2}$. If $val(c)$ indicates the local resistance value

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 6.22: Computation of diffusion on a raster. The lower left green cell is the source location, indicated in the local resistance raster (a). The raster in (b) is the minimal total resistance raster computed by the GIS. (The GIS will work in higher precision real arithmetic than what is illustrated here.)

for cell c , the GIS computes the total incurred resistance for diffusion from c_{src} to c_e as $\frac{1}{2}(val(c_{\text{src}}) + val(c_e))$, while the same for c_{src} to c_{ne} is $\frac{1}{2}(val(c_{\text{src}}) + val(c_{\text{ne}})) \times \sqrt{2}$. 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 6.22(b).

Minimal cost path

For instance, there are three paths from the green source location to its northeast neighbour cell (with local resistance 5). We can define them as path 1 (N–E),

path 2 (E–N) and path 3 (NE), using compass directions to define the path from the green cell. For path 1, the total resistance is computed as:

$$\frac{1}{2}(4 + 4) + \frac{1}{2}(4 + 5) = 8.5.$$

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.

6.4.3 Flow computation

Flow computations determine how a phenomenon spreads over the area, in principle in all directions, though with varying difficulty or resistance. There are also cases where a phenomenon does not spread in all directions, but moves or ‘flows’ along a given, 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.

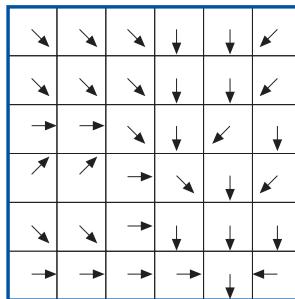
This principle is illustrated with a simple elevation raster, in Figure 6.23(a). For each cell in that raster, the steepest downward slope to a neighbour cell is computed, and its direction is stored in a new raster (Figure 6.23(b)). This computation determines the elevation difference between the cell 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.

Determining flow direction

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 map algebra 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.

156	144	138	142	116	98
148	134	112	98	92	100
138	106	88	74	76	96
128	116	110	44	62	48
136	122	94	42	32	38
148	106	68	24	22	24

(a)



(b)

0	0	0	0	0	0
0	1	1	2	2	0
0	3	7	5	4	0
0	0	0	20	0	1
0	0	0	1	24	0
0	2	4	7	35	1

(c)

Figure 6.23: Flow computations on a raster: (a) the original elevation raster, (b) the flow direction raster computed from it, (c) accumulated flow count raster.

6.4.4 Raster based surface analysis

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

The discussions in this section 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 regularly refer to the continuous field representation as a DEM, to conform with the most common situation. Throughout the section we will assume that the DEM is represented as a raster.

Applications

There are numerous examples where more advanced computations on continuous field representations are needed. A short list is provided below.

- *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. Filters are discussed on page 6.4.4. 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 and lava flows.

- *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 computations mentioned above are further discussed below. All of them apply a technique known as *filtering*, so we will first examine this principle in more detail.

Filtering

The principle of filtering is quite similar to that of *moving window averaging*, which we discussed in Section 5.4.2. 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.³ The difference with moving window averaging is that the moving window in filtering is itself a little raster, which contains cell values that are used in the computation for the output cell value. This little raster is a *filter*, also known as a *kernel* which may be square (such as a 3x3 kernel), but it does not have to be. The values in the filter are used as weight factors.

Window or kernel

As an example, let us consider a 3×3 cell filter, in which all values are equal to 1, as illustrated in Figure 6.24(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 that we divide by the sum of *absolute* weights.

³Please refer to Chapter Five of *Principles of Remote Sensing* for a discussion of image-related filter operations.

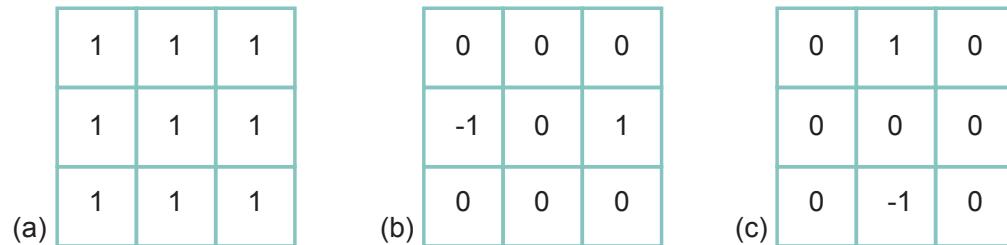


Figure 6.24: Moving window rasters for filtering.
(a) raster for a regular averaging filter; (b) raster for an x -gradient filter;
(c) raster for a y -gradient filter.

Since the w_{ij} are all equal to 1 in the case of Figure 6.24(a), the formula can be simplified to

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

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, so its application amounts to moving window averaging. 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.

Computation of slope angle and slope aspect

A different choice of weight factors may provide other information. Special filters exist to perform computations on the slope of the terrain. Before we look at these filters, let us define various notions of *slope*.

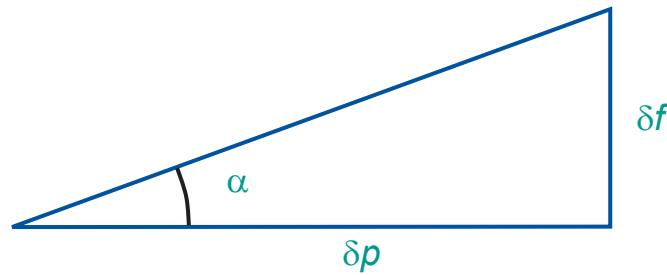


Figure 6.25: Slope angle defined. Here, δp stands for length in the horizontal plane, δf stands for the change in field value, where the field usually is terrain elevation. The slope angle is α .

Slope angle, which is also known as *slope gradient*, is the angle α , illustrated in Figure 6.25, 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:

$$\text{slope_perc} = 100 \cdot \frac{\delta f}{\delta p} \text{ and } \text{slope_angle} = \arctan\left(\frac{\delta f}{\delta p}\right).$$

The path p must be chosen to provide the highest slope angle value, and thus it can lie in any direction. The compass direction, converted to an angle 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 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 6.24(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.

x and *y* gradient filters

Observe that both filters express elevation gain *per cell width*. This means that we must divide by the cell width—given in metres, for example—to obtain the (approximations to) the true derivatives $\delta f/\delta x$ and $\delta f/\delta y$. Here, f stands for the elevation field as a function of x and y , and $\delta f/\delta x$, for instance, is the elevation gain per unit of length in the x -direction.

To obtain the real slope angle α along path p , observe that both the x - and y -gradient contribute to it. This is illustrated in Figure 6.26. 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 6.24(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 $\sqrt{\cdot \cdot \cdot}$ 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

$$\tan(\psi) = \frac{\delta f / \delta x}{\delta f / \delta y},$$

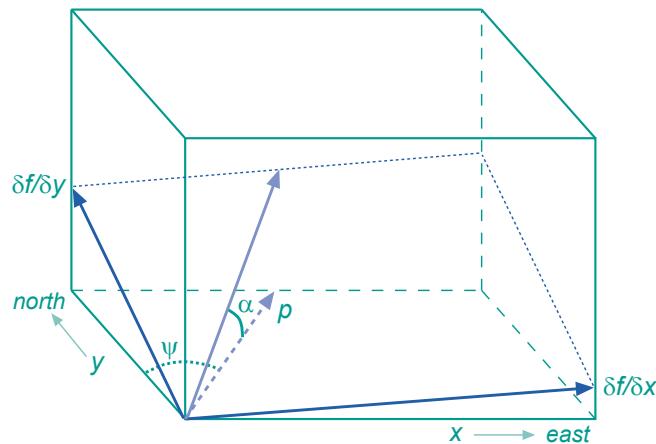


Figure 6.26: Slope angle and slope aspect defined. Here, p is the horizontal path in maximal slope direction and α is the slope angle. The plane tangent to the terrain in the origin is also indicated. The angle ψ is the slope aspect. See the text for further explanation.

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

$$\psi = \arctan\left(\frac{\delta f/\delta x}{\delta f/\delta y}\right),$$

the reason being that the latter formula does not account for southeast and southwest quadrants, nor for cases where $\delta f/\delta y = 0$. (In the first situation, one must add 180° to the computed angle to obtain an angle measured from North; in the latter situation, ψ equals either 90° or -90° , depending on the sign of $\delta f/\delta x$.)

6.5 Network analysis

A completely different set of analytical 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 performed on either raster or vector data layers, but they are more commonly done in the latter, as line features can be associated with a network, and hence can be assigned typical transportation characteristics such as *capacity* and *cost per unit*. A fundamental characteristic of any network is whether the network lines are considered directed or not. *Directed networks* associate with each line a direction of transportation; *undirected networks* do not. In the latter, the ‘goods’ can be transported along a line in both directions. We discuss here vector network analysis, and assume that the network is a set of connected line features that intersect only at the lines’ nodes, not at internal vertices. (But we do mention under- and overpasses.)

Directed and undirected networks

For many applications of network analysis, a *planar network*, i.e. one that can be embedded in a two-dimensional plane, will do the job. Many networks are naturally planar, 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.

Planar networks

Not all GISs accommodate non-planar networks, or can do so only using ‘tricks’.

These may involve the splitting of overpassing lines at the intersection vertex and the creation of four lines out of the two original lines. Without further attention, the network will then allow one to make a turn onto another line at this new intersection node, which in reality would be impossible. In some GISs we can allocate a cost with turning at a node—see our discussion on turning costs below—and that cost, in the case of the overpass, can be made infinite to ensure it is prohibited. But, as mentioned, this is a workaround to fit a non-planar situation into a data layer that presumes planarity.

Overpasses

The above is a good illustration of geometry not fully determining the network's behaviour. Additional application-specific rules are usually required to define what can and cannot happen in the network. Most GISs provide rule-based tools that allow the definition of these extra application rules.

Various classical spatial analysis functions on networks are supported by GIS software packages. The most important ones are:

1. *Optimal path finding* which generates a least cost-path on a network between a pair of predefined locations using both geometric and attribute data.
2. *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 6.27.

Turning costs

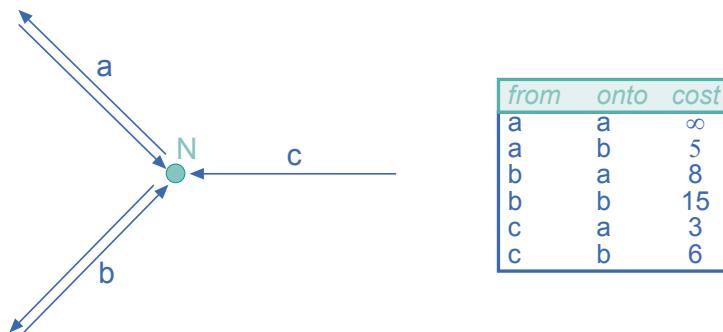


Figure 6.27: Network neighbourhood of node N with associated turning costs at N . Turning at N onto c is prohibited because of direction, so no costs are mentioned for turning onto c . A turning cost of infinity (∞) means that it is also prohibited.

The attentive reader will notice that it is possible to travel on line b in Figure 6.27,

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 an 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 6.28. 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 GISs provide support for these more complicated path finding problems.

Ordered and unordered path
finding

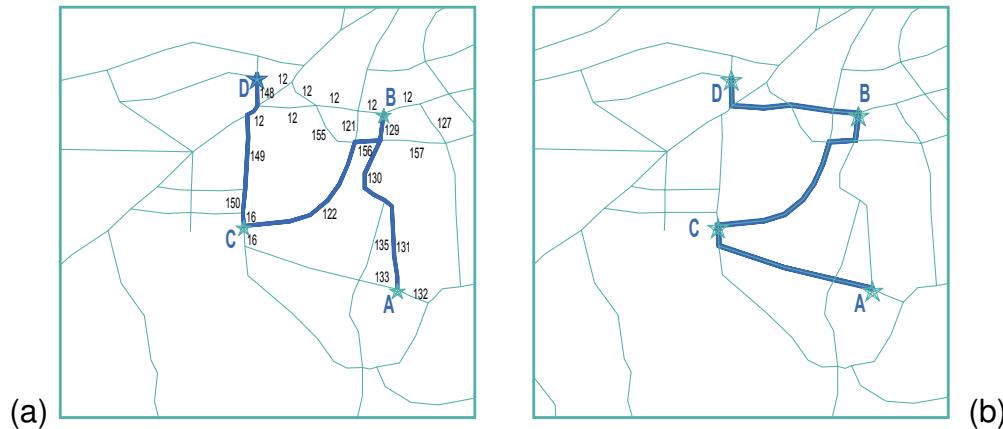


Figure 6.28: Ordered (a) and unordered (b) optimal path finding. In both cases, a path had to be found from A to D , in (a) by visiting B and then C , in (b) both nodes also but in arbitrary order.

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*.

Service areas

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.

Connectivity

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

- The *capacity* with which a centre can produce the resources (whether they are medical operations, school pupil positions, kilowatts, or bottles of milk), and
- The *consumption of the resources*, which may vary amongst lines or line segments. After all, some streets have more accidents, more children who

live there, more industry in high demand of electricity or just more thirsty workers.

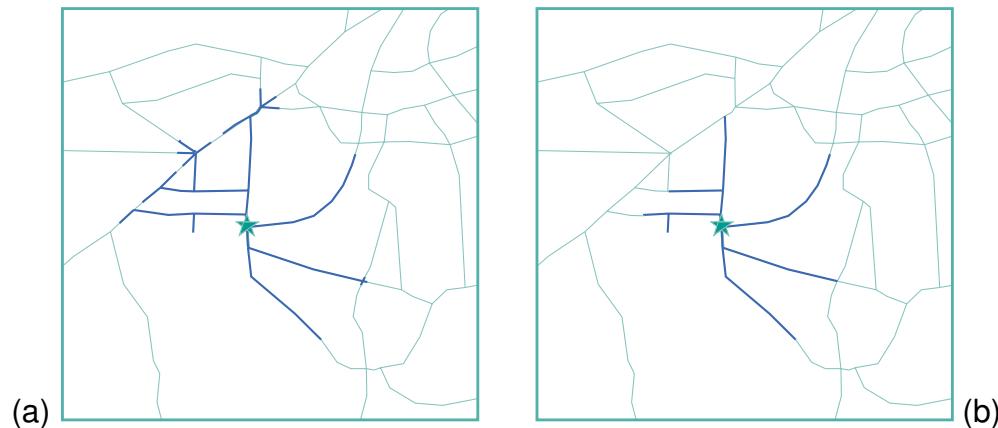


Figure 6.29: Network allocation on a pupil/school assignment problem. In (a), the street segments within 2 km of the school are identified; in (b), the selection of (a) is further restricted to accommodate the school's capacity for the new year.

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 6.29(a), the green star indicates a primary school and the GIS has been used to assign streets and street segments 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, or resource consumption along it. The condition typically is a logical expression, as we have seen before, for instance:

Tracing requires connectivity

- 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.

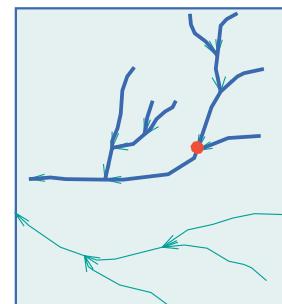
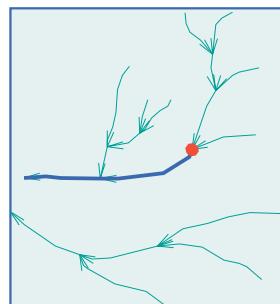
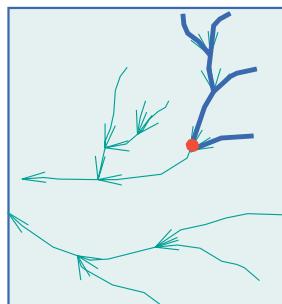


Figure 6.30: Tracing functions on a network:
(a) tracing upstream,
(b) tracing downstream,
(c) tracing without conditions on direction.

In Figure 6.30 our trace origin is indicated in red. In part (a), the tracing conditions were set to trace all the way upstream; part (b) traces all the way downstream, and in part (c) there are no conditions on direction of the path, thereby tracing all connected lines from the trace origin. More complex conditions are certainly possible in tracing.

Upstream and downstream
tracing

6.6 GIS and application models

We have discussed the notion that real world processes are often highly complex. Models are simplified abstractions of reality representing or describing its most important elements and their interactions. Modelling and GIS are more or less inseparable, as GIS is itself a tool for modelling ‘the real world’ (or at least some part of it).

The solution to a (spatial) problem usually depends on a (large) number of parameters. Since these parameters are often interrelated, their interaction is made more precise in an *application model*.

Here we define application models to include any kind of GIS based model (including so-called analytical and process models) for a specific real-world application. Such a model, in one way or other, describes as faithfully as possible how the relevant geographic phenomena behave, and it does so in terms of the parameters.

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 to derive a solution. Many kinds of application models exist, and they can be classified in many different ways. Here we identify five characteristics of GIS-based application models:

1. The *purpose* of the model,
2. The *methodology* underlying the model,
3. The *scale* at which the model works,

4. Its *dimensionality* - i.e. whether the model includes spatial, temporal or spatial and temporal dimensions, and
5. Its *implementation logic* - i.e. the extent to which the model uses existing knowledge about the implementation context.

Model characteristics

It is important to note that the categories above are merely *different characteristics* of any given application model. Any model can be described according to these characteristics. Each is briefly discussed below.

Purpose of the model refers to whether the model is descriptive, prescriptive or predictive in nature. *Descriptive models* attempt to answer the “what is” - question. *Prescriptive models* usually answer the “what should be” question by determining the best solution from a given set of conditions.

Descriptive and prescriptive models

Models for planning and site selection are usually prescriptive, in that they quantify environmental, economic and social factors to determine ‘best’ or optimal locations. So-called *Predictive models* focus upon the “what is likely to be” questions, and predict outcomes based upon a set of input conditions. Examples of predictive models include forecasting models, such as those attempting to predict landslides or sea-level rise.

Predictive models

Methodology refers to the operational components of the model. *Stochastic* models use statistical or probability functions to represent random or semi-random behaviour of phenomena. In contrast, *deterministic* models are based upon a well-defined cause and effect relationship. Examples of deterministic models

Inner workings of the model

include hydrological flow and pollution models, where the ‘effect’ can often be described by numerical methods and differential equations.

Rule-based models attempt to model processes by using local (spatial) rules. *Cellular Automata* (CA) are examples of models in this category. These are often used to understand systems which are generally not well understood, but for which their local processes are well known. For example, the characteristics of neighbourhood cells (such as wind direction and vegetation type) in a raster-based CA model might be used to model the direction of spread of a fire over several time steps.

Agent-based models (ABM) attempt to model movement and development of multiple interacting agents (which might represent individuals), often using sets of decision-rules about what the agent can and cannot do. Complex agent-based models have been developed to understand aspects of travel behaviour and crowd interactions which also incorporate stochastic components.

Scale refers to whether the components of the model are *individual* or *aggregate* in nature. Essentially this refers to the ‘level’ at which the model operates. *Individual-based* models are based on individual entities, such as the agent-based models described above, whereas *aggregate* models deal with ‘grouped’ data, such as population census data. Aggregate models may operate on data at the level of a city block (for example, using population census data for particular social groups), at the regional, or even at a global scale.

Individual and aggregate
models

Dimensionality is the term chosen to refer to whether a model is static or dynamic, and spatial or aspatial. Some models are explicitly spatial, meaning they

operate in some geographically defined space. Some models are *aspatial*, meaning they have no direct spatial reference.

Models can also be *static*, meaning they do not incorporate a notion of time or change. In *dynamic* models, time is an essential parameter (see Section 2.5). Dynamic models include various types of models referred to as process models or simulations. These types of models aim to generate future scenarios from existing scenarios, and might include deterministic or stochastic components, or some kind of local rule (for example, to drive a simulation of urban growth and spread). The fire spread example given above is a good example of an explicitly spatial, dynamic model which might incorporate both local rules and stochastic components.

Static and dynamic models

Implementation logic refers to how the model uses existing theory or knowledge to create new knowledge. *Deductive* approaches use knowledge of the overall situation in order to predict outcome conditions. This includes models that have some kind of formalized set of criteria, often with known weightings for the inputs, and existing algorithms are used to derive outcomes. *Inductive* approaches, on the other hand, are less straightforward, in that they try to generalize (often based upon samples of a specific data set) in order to derive more general models. While an inductive approach is useful if we do not know the general conditions or rules which apply in a given domain, it is typically a trial-and-error approach which requires empirical testing to determine the parameters of each input variable.

Inductive and deductive approaches

Most GIS only come equipped with a limited range of tools for modelling. For complex models, or functions which are not natively supported in our GIS, exter-

nal software environments are frequently used. In some cases, GIS and models can be fully integrated (known as *embedded coupling*) or linked through data and interface (known as *tight coupling*). If neither of these is possible, the external model might be run independently of our GIS, and the output exported from our model into the GIS for further analysis and visualization. This is known as *loose coupling*.

Coupling

It is important to compare our model results with previous experiments and to examine the possible causes of inconsistency between the output of our models and the expected results. The following section discusses these aspects further.

6.7 Error propagation in spatial data processing

6.7.1 How errors propagate

In Section 5.2, we discussed a number of sources of error that may be present in source data. It is important to note that the acquisition of base data to a high standard of quality still does not guarantee that the results of further, complex processing can be treated with certainty. As the number of processing steps increases, it becomes difficult to predict the behaviour of *error propagation*. These various errors may affect the outcome of spatial data manipulations. In addition, further errors may be introduced during the various processing steps discussed earlier in this chapter, as illustrated in Figure 6.31.

Combined error from individual sources

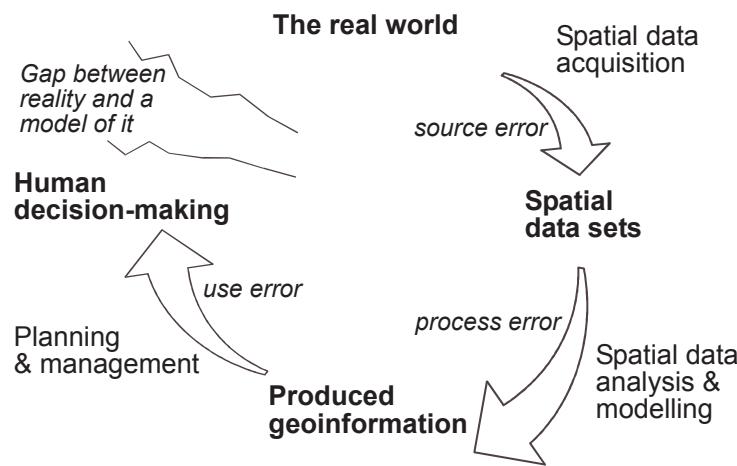


Figure 6.31: Error propagation in spatial data handling

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.

Table 6.2 lists common sources of error introduced into GIS analyses. Note that these are from a wide range of sources, and include various common tasks relating to both data preparation and data analysis. It is the combination of different errors that are generated at each stage of preparation and analysis which may bring about various errors and uncertainties in the eventual outputs.

Consider another example. A land use planning agency is faced with the problem of identifying areas of agricultural land that are highly susceptible to erosion. Such areas occur on steep slopes in areas of high rainfall. The spatial data used in a GIS to obtain this information might include:

- A land use map produced five years previously from 1 : 25,000 scale aerial photographs,
- A DEM produced by interpolating contours from a 1 : 50,000 scale topographic map, and
- Annual rainfall statistics collected at two rainfall gauges.

The reader is invited to assess what sort of errors are likely to occur in this analysis.

Referring back to Figure 6.31, the reader is also encouraged to reflect on errors introduced in components of application models discussed in the previous section. Specifically, the methodological aspects of representing geographic phenomena. What might be the consequences of using a random function in an urban transportation model (when, in fact, travel behaviour is not purely random)?

<i>Coordinate adjustments</i>	<i>Generalization</i>
rubber sheeting/transformations	linear alignment
projection changes	line simplification
datum conversions	addition/deletion of vertices
rescaling	linear displacement
<i>Feature Editing</i>	<i>Raster/Vector Conversions</i>
line snapping	raster cells to polygons
extension of lines to intersection	polygons to raster cells
reshaping	assignment of point attributes
moving/copying	to raster cells
elimination of spurious polygons	post-scanner line thinning
<i>Attribute editing</i>	<i>Data input and Management</i>
numeric calculation and change	digitizing
text value changes/substitution	scanning
re-definition of attributes	topological construction / spatial indexing
attribute value update	dissolving polygons with same attributes
<i>Boolean Operations</i>	<i>Surface modelling</i>
polygon on polygon	contour/lattice generation
polygon on line	TIN formation
polygon on point	Draping of data sets
line on line	Cross-section/profile generation
overlay and erase/update	Slope/aspect determination
<i>Display and Analysis</i>	<i>Display and Analysis</i>
cluster analysis	class intervals choice
calculation of surface lengths	areal interpolation
shortest route/path computation	perimeter/area size/volume computation
buffer creation	distance computation
display and query	spatial statistics
adjacency/contiguity	label/text placement

Table 6.2: Some of the most common causes of error in spatial data handling. Source: Hunter & Beard [23].

6.7.2 Quantifying error propagation

Chrismen [13] noted that “the ultimate arbiter of cartographic error is the real world, not a mathematical formulation”. It is an unavoidable fact that we will never be able to capture and represent everything that happens in the real world perfectly in a GIS. Hence there is much to recommend the use of testing procedures for accuracy assessment.

Errors are unavoidable

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. All these approaches have their origins in academic research and have strong theoretical bases in mathematics and statistics. Here we identify two main approaches for assessing the nature and amount of error propagation:

1. Testing the accuracy of each state by *measurement against the real world*, and
2. *Modelling* error propagation, either analytically or by means of simulation techniques.

Modelling of error propagation has been defined by Veregin [56] as: “the application of formal mathematical models that describe the mechanisms whereby errors in source data layers are modified by particular data transformation operations.” In other words, we would like to know how errors in the source data behave under manipulations that we subject them to in a GIS. If we are able to quantify the error in the source data as well as their behaviour under GIS manipulations, we have a means of judging the uncertainty of the results.

Modelling error vs.
modelling error propagation

Error propagation models are very complex and valid only for certain data types (e.g. numerical attributes). Initially, they described only the propagation of attribute error [21, 56]. More recent research has addressed the spatial aspects of error propagation and the development of models incorporating both attribute and locational components. These topics are outside the scope of this book, and readers are referred to [2, 27] for more detailed discussions. Rather than explicitly modelling error propagation, it is often more practical to test the results of each step in the process against some independently measured reference data.

Attribute and locational
components

Summary

This chapter has examined various ways of manipulating both raster and vector based spatial data sets. It is certainly true that some types of manipulations are better accommodated in one, and not so well in the other. Usually, one chooses the format to work with on the basis of many more parameters, including the availability of source data.

We have identified several *classes* of data manipulations or functions. The first of these 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. *Spatial selections* allow us to selectively identify features on the basis of conditions, which may be spatial in character.

A second class of spatial data manipulations generates new spatial data sets. *Classification* functions assign a new characteristic value to each feature in a set of (previously selected) features. *Spatial overlay* functions go a step further and combine two spatial data sets by location. What is produced as an 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 context, and therefore are sensitive to direction. Since such local factors are more easily represented in raster data, this is then the preferred format. Flow and diffusion functions are examples.

We also looked at a special type of spatial data, namely (line) *networks*, and the functions that are used on these. *Optimal path finding* is one such function, useful in routing problems. The use of this function can be constrained or unconstrained. Another function often used on networks is *network partitioning*: how to assign respective parts of the network to resource locations.

Various combinations of the analytical functions discussed above can be used in an *application model* to simulate a given geographical process or phenomenon. The output generated by these models can then be used in various ways, including decision support and planning. Many different kinds of models exist, and the type of model used will depend on the process or phenomena under study, the nature of the data, and the type of output desired from the model.

The final section of this chapter discussed the issue of error propagation. It was noted that at each stage of working with spatial data, errors can be introduced which can *propagate* through the different operations. These errors can range from simple mistakes in data entry through to inappropriate estimation techniques or functions in operational models, and can serve to degrade the 'end result' of our analyses significantly.

Questions

1. On page 352, 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 352, 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 6.11 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. In Figure 6.9, we provided a classification of average household income per ward in the city of Dar es Salaam. Provide a (spatial) interpretation of that figure.
5. Observe that the *equal frequency technique* applied on the raster of Figure 6.11 does not really produce categories with equal frequencies. Explain why this is. Would we expect a better result if our raster had been $5,000 \times 5,000$ cells?

6. 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.
7. Argue why diffusion computations are much more naturally supported by raster data than by vector data.
8. In Figure 6.22(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.
9. In Figure 6.23, 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?
10. In Section 6.4.4, 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.



Chapter 7

Data visualization

7.1 GIS and maps

There is a strong relationship between maps and GIS. More specifically, maps can be used as input for a GIS. They play a key role in relation to all the functional components of a GIS shown in Figure 3.1.

As soon as a question contains a “where?” question, a map can often 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, they do not give the full picture.

“Where?”

A map would put these answers in a spatial context. It could show where in the Netherlands Enschede is to be found and where it is located with respect to Schiphol–Amsterdam airport, where most students arrive. 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 7.1.

As soon as the location of geographic objects (“where?”) is involved, a map becomes useful. However, maps can do more than 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 7.2, a dominant northwest-southeast urban buffer can be clearly distinguished. Maps can answer the “What?” question only in relation to location (the map as a reference

“What?”

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 7.3.

“When?”

To summarize, maps can deal with questions/answers related to the basic components of spatial or geographic data: location (geometry), characteristics (themetic attributes) and time, and their combination.

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

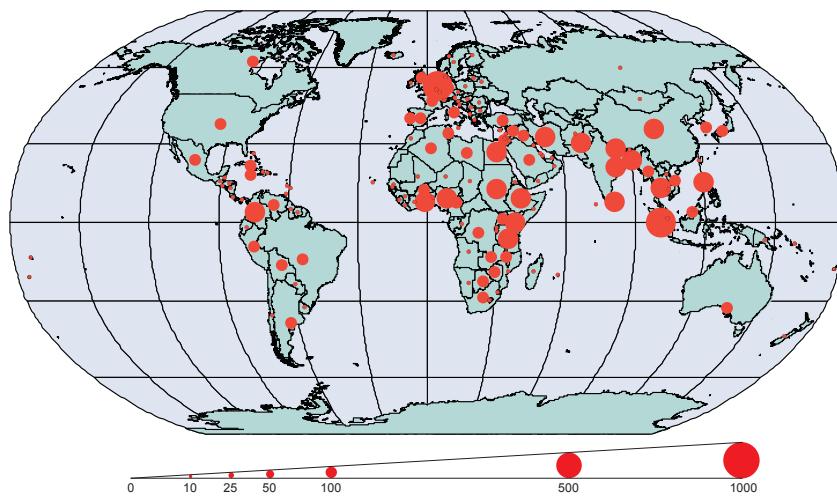


Figure 7.1: Maps and location—“Where did ITC cartography students come from?” Map scale is 1 : 200,000,000.

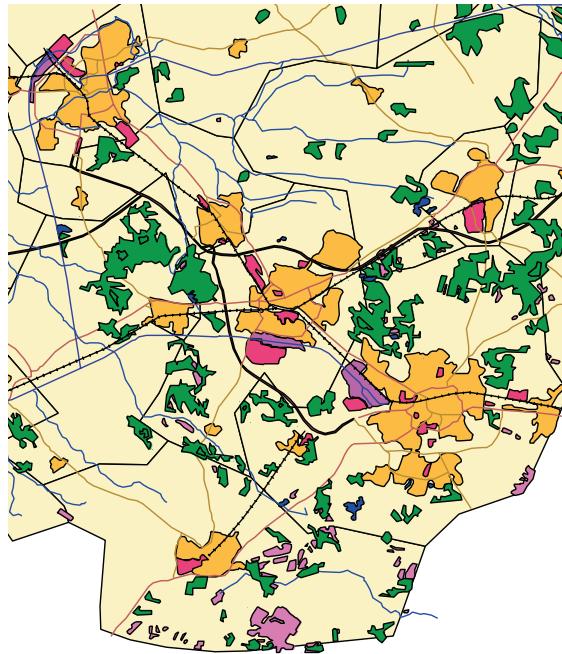


Figure 7.2: Maps and characteristics—“What is the predominant land use in southeast Twente?”

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 above 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 infor-

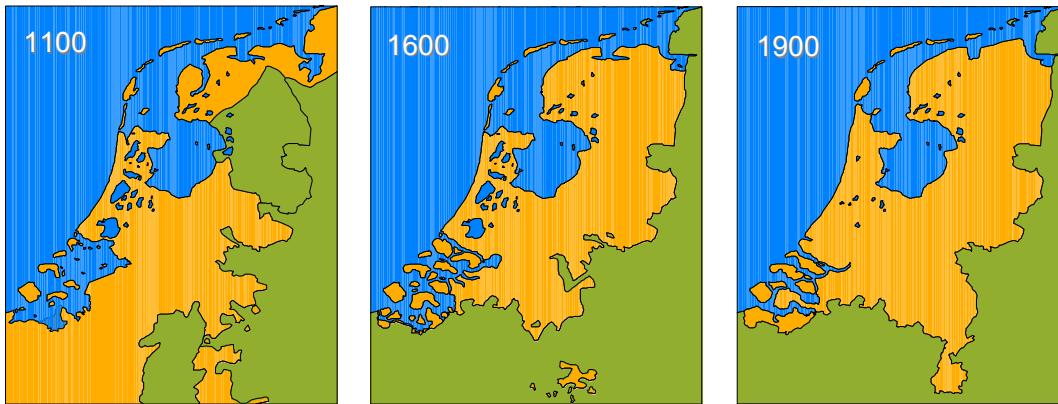


Figure 7.3: Maps and time—“When did the Netherlands have its longest coastline?”

mation in a clear perspective. The map in Figure 7.1 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.

Simplification and abstraction from reality

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 7.4 shows an aerial photograph of the ITC building and a map of the same area. The photographs show all visible objects, including parked cars, and small temporary buildings. From the photograph, it becomes clear that the weather as well as the time of the day influenced its contents: the shadow to the north of the buildings obscures other information. The map on the other hand, only gives the outlines of buildings and the streets in the surroundings. It is easier to interpret because of selection/omission and classification of features. 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.

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 7.1 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 7.4 displaying the surroundings of the ITC-building is an example. The world map in Figure 7.1 is a *small-scale map*. Scale indications on maps can be given verbally

Map scale

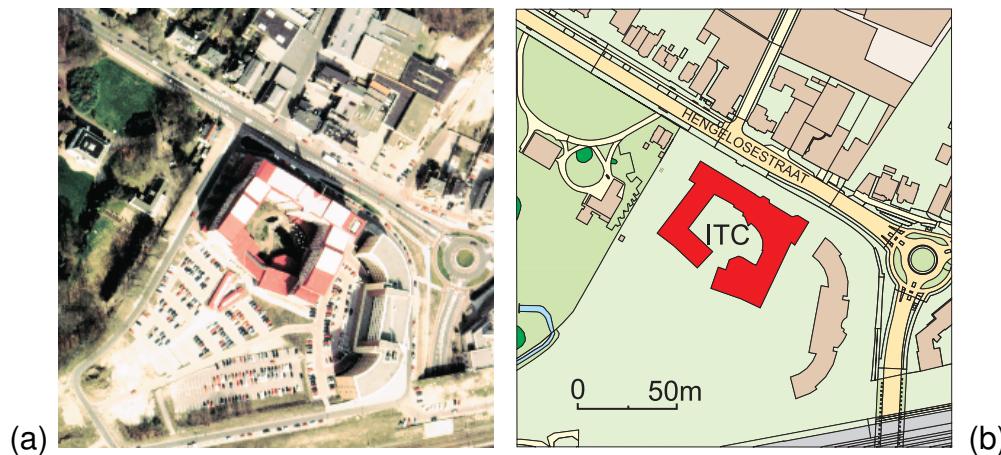


Figure 7.4: Comparing an aerial photograph (a) and a map (b). Source: Figure 5–1 in [30].

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 7.4(b). The advantage of using scale bars in digital environments is that its length changes also when the map is 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 provide a definition. Board [7] 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 into *topographic maps* and *thematic maps*. A topographic map visualizes, limited by its scale, the Earth’s surface as accurately as possible. This may include infrastructure (e.g. railroads and roads), land use (e.g. vegetation and built-up area), relief, hydrology, geographic names and a

Topographic maps

¹This explains why many of the maps in this book do not show a map scale.

reference grid. Figure 7.5 shows a small scale topographic map (with text omitted) of Overijssel, the Dutch province in which Enschede is located.

Thematic maps represent the distribution of particular *themes*. One can distinguish between *socio-economic themes* and *physical themes*. The map in Figure 7.6(a), showing population density in Overijssel, is an example of the first and the map in Figure 7.6(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 well.

Thematic maps

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.

The design of topographic maps is mostly based on conventions, of which some date back several centuries. Examples are the use of blue to represent water, green for forests, red for major roads, and black to denote urban or built-up areas. The design of thematic maps, however, should be based on a set of cartographic rules, also called *cartographic grammar*, which will be explained in Sections 7.4 and 7.5 (but see also [32]).

Cartographic grammar

Suppose that one wants to quantify land use changes between 1990 and the current year. Two data sets (from 1990 and 2008) can be combined with an overlay

operation (see Section 6.3). 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 models developed by the application at hand. It is easy to imagine that maps can play a role during this process of working with a GIS by showing intermediate and final results of the GIS operations. Clearly, maps are no longer only the final product they used to be.

Maps can further be distinguished according to the dimensions of spatial data that are graphically represented. GIS users also try to solve problems that deal with three-dimensional reality or with change processes. This results in a demand for other than just two-dimensional maps to represent geographic reality. Three-dimensional and even four-dimensional (namely, including time) maps are then required. New visualization techniques for these demands have been developed. Figure 7.7 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 7.7(c) shows the effect of change, as three moments in time during the construction of the building.

Dimensionality

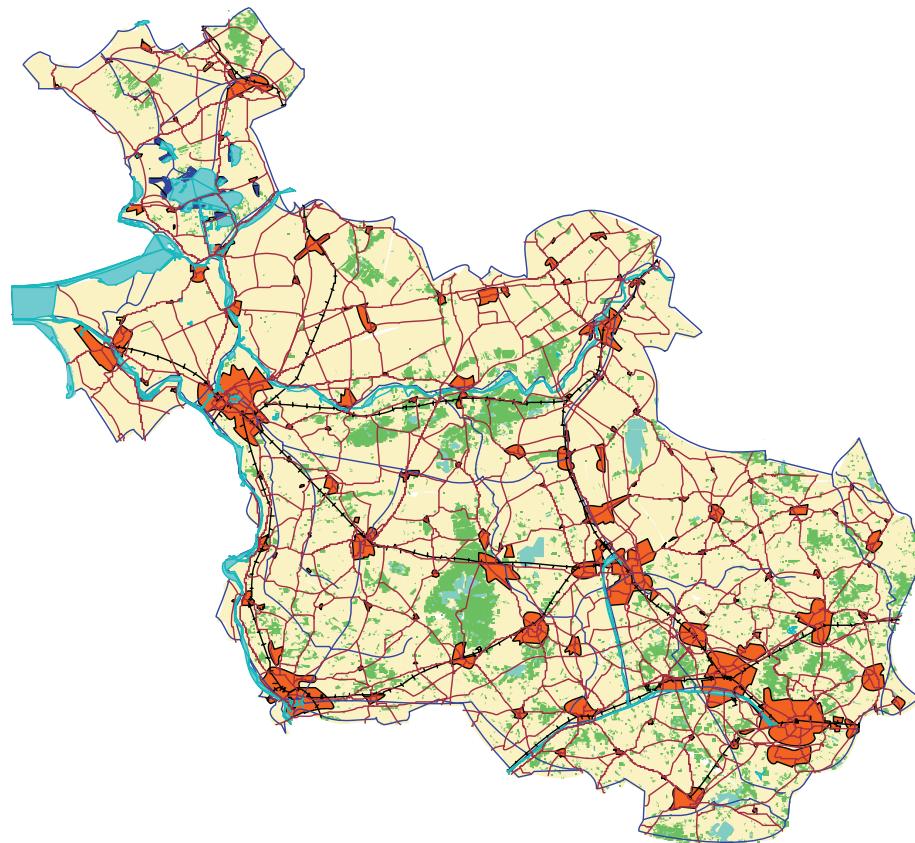


Figure 7.5: A topographic map of the province of Overijssel. Geographic names and a reference grid have been omitted for reasons of clarity.

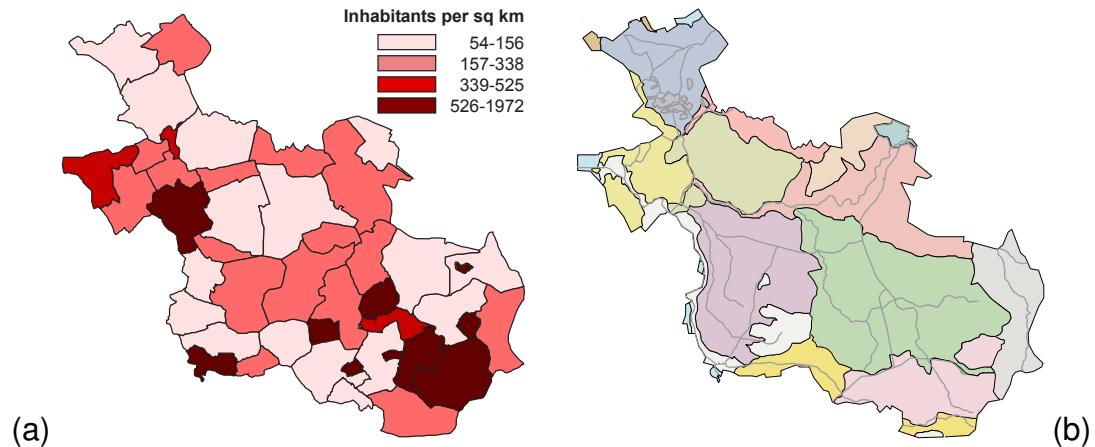


Figure 7.6: Thematic maps: (a) socio-economic thematic map, showing population density of the province of Overijssel (higher densities in darker tints); (b) physical thematic map, showing watershed areas of Overijssel.

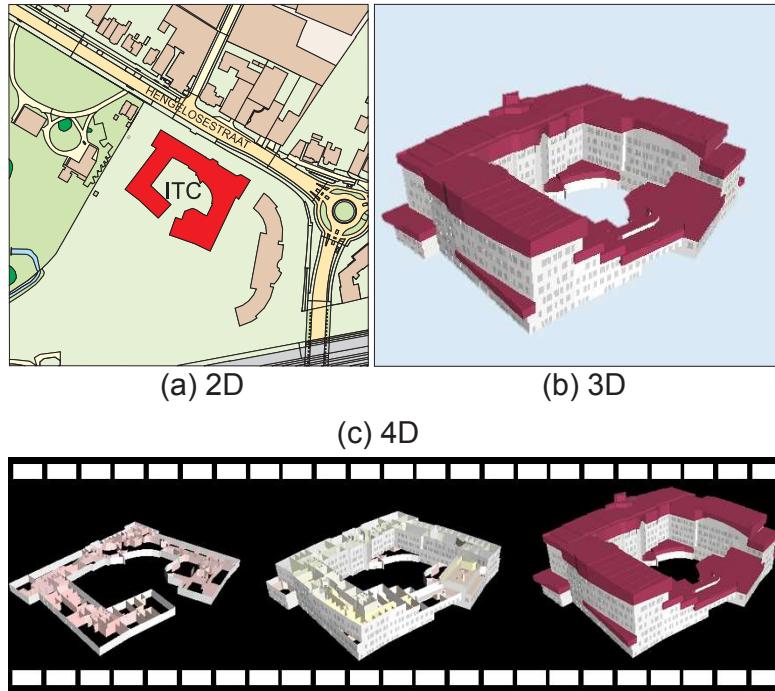


Figure 7.7: The dimensions of spatial data: (a) 2D, (b) 3D, (c) 3D with time.

7.2 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 and production for the use of maps, depending on the application (see Figure 7.8).

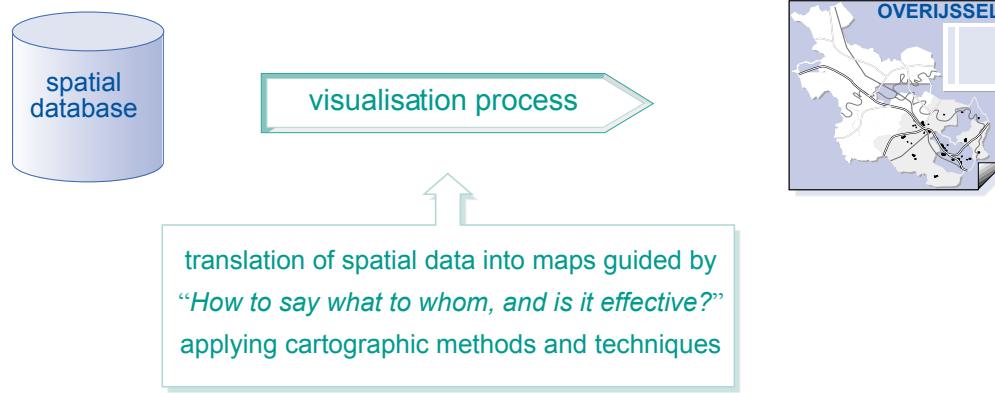


Figure 7.8: The cartographic visualization process. Source: Figure 2–1 in [30].

The producer of these visual products may be a professional cartographer, but may also be a discipline expert, for instance, mapping 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).

Visualization purpose and environment

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 effectiveness. To the proposition “How do I say what to whom” we have to add “and is it effective?” Based on feedback from map users, or knowledge about the effectiveness of cartographic solutions, we can decide whether improvements are needed, and derive recommendations for future application of those solutions. In particular, with all the visualization options available, such as animated maps, multimedia and virtual reality, it remains necessary to test the effectiveness of cartographic methods and tools.

Effectiveness

The visualization process is always influenced by several factors. Some of these questions can be answered by just looking at the content of the spatial database:

- What will be the scale of the map: large, small, other? This introduces the problem of generalization. *Generalization* addresses the meaningful reduction of the map content during scale reduction.
- Are we dealing with topographic or thematic data? These two categories 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 increase, since the compilation of maps by spatial data handling is often the result of combining different data sets of different quality and from different data sources, collected at different scales and stored in different map projections.

Cartographers have all kind of tools available to visualize the data. These tools consist of functions, rules and habits. Algorithms used to classify the data or to smooth a polyline are examples of functions. Rules tell us, for instance, to use proportional symbols to display absolute quantities or to position an artificial light source in the northwest to create a shaded relief map. Habits or conventions—or traditions as some would call them—tell us to colour the sea 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.

Cartographic rules

7.3 Visualization strategies: present or explore?

Traditionally the cartographer's main task was the creation of good cartographic products. This is still true today. The main function of maps is to communicate geographic information, i.e. to inform the map user about location and 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 [50, 32].

Visual communication

During the last decades, many others have become involved in making maps. The widespread use of GIS has increased the number of maps tremendously [35]. 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.

Visual thinking process

This thinking process is accelerated by the continued developments in hardware and software. Media like DVD-ROMs and the WWW allow *dynamic presentation* and also *user interaction*. These went along with changing scientific and societal needs for georeferenced data and, as such, for maps. Users now expect immediate and real-time access to the data; data that have become abundant in many sectors of the geoinformation world. This abundance of data, seen as a 'paradise' by some sectors, is a major problem in other sectors. We lack the tools for user-friendly queries and retrieval when studying the massive amount of (spatial) data produced by sensors, which is now available via the WWW. A new

Visual data mining

branch of science is currently evolving to deal with this problem of abundance. In the geo-disciplines, it is called *visual data mining*.

These developments have given the term *visualization* an enhanced meaning. According to the dictionary, it means 'to make visible' or 'to represent in graphical form'. It can be argued that, in the case of spatial data, this has always been the business of cartographers. However, progress in other disciplines has linked the word to more specific ways in which modern computer technology can facilitate the process of 'making visible' in real time. Specific software toolboxes have been developed, and their functionality is based on two key words: *interaction* and *dynamics*. A separate discipline, called scientific visualization, has developed around it [37], and has also had an important impact on cartography. 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.

Interaction and dynamics

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 7.9). 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

Geovisualization

expert's attempt to solve a problem. While dealing with the data, the expert should be able to rely on cartographic expertise, provided by the software or some other means. Essentially, here the problem of translation of spatial data into cartographic symbols also needs to be solved.

The above trends all have to do with what has been called the 'democratization of cartography' by Morrison [40]. He explains it as

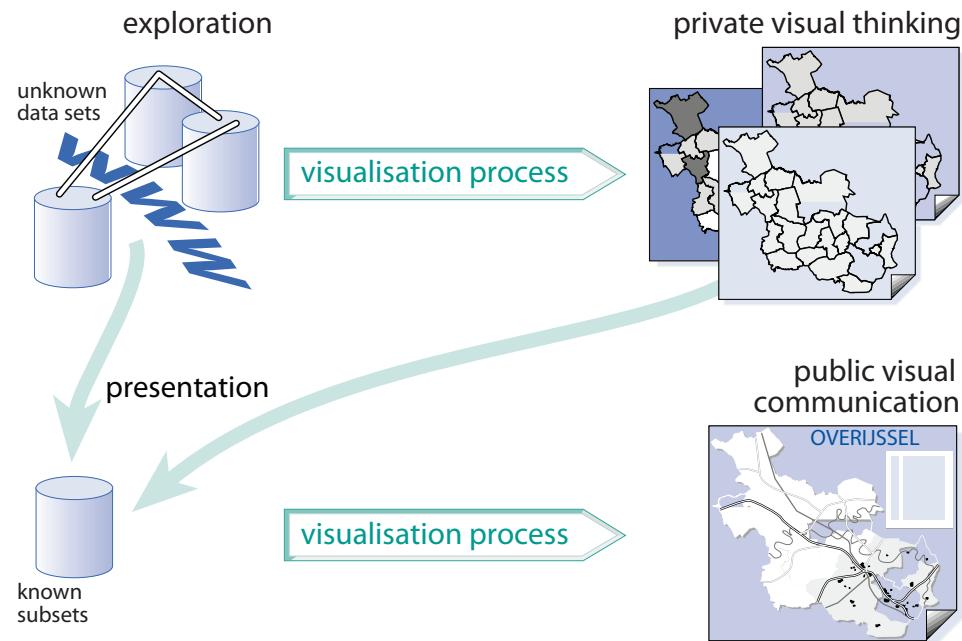


Figure 7.9: Private visual thinking and public visual communication. Source: Modified from Figure 2–2 in [30].

“using electronic technology, no longer does the map user depend on what the cartographer decides to put on a map. Today the user is the cartographer ... users are now able to produce analyses and visualizations at will to any accuracy standard that satisfies them.”

Exploration means to search for spatial, temporal or spatio-temporal patterns, relationships between patterns, or trends. In case of a search for patterns, a domain expert may be interested in aspects like the distribution of a phenomenon, the occurrence of anomalies, the sequence of appearances and disappearances. A search for relationships between patterns could include: changes in vegetation indices and climatic parameters, location of deprived urban areas and their distance to educational facilities. A search for trends could, for example, focus on the development in distribution and frequency of landslides. Maps not only enable these types of searches, findings may also trigger new questions, and lead to new visual exploration (or analytical) acts.

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 the same users will probably no longer feel constrained as soon as they follow the web links attached to this electronic atlas. This shows that the data, the users, and the use environment influence one’s view of what exploration entails.

To create a map, one selects relevant geographic data and converts these into meaningful symbols for the map. 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 in the realm of presentation cartography (Figure 7.9), it is important to adhere to the cartographic design rules. This is to guarantee that the resulting maps are easily understood by their users. How does this communication process work? Figure 7.10 forms an illustration. It starts with information to be mapped (the ‘What’ from the sentence). Before anything can be done, the cartographer should get a feel for the nature of the information, since this determines the graphical options. Cartographic information analysis provides 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 7.4.

Visual communication
process

In 1967, the French cartographer Bertin developed the basic concepts of the theory of map design, with his publication *Sémiologie Graphique* [5]. He provided guidelines for making good maps. If ten professional cartographers were given the same mapping task, and each would apply Bertin’s rules (see Section 7.4.2), 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.

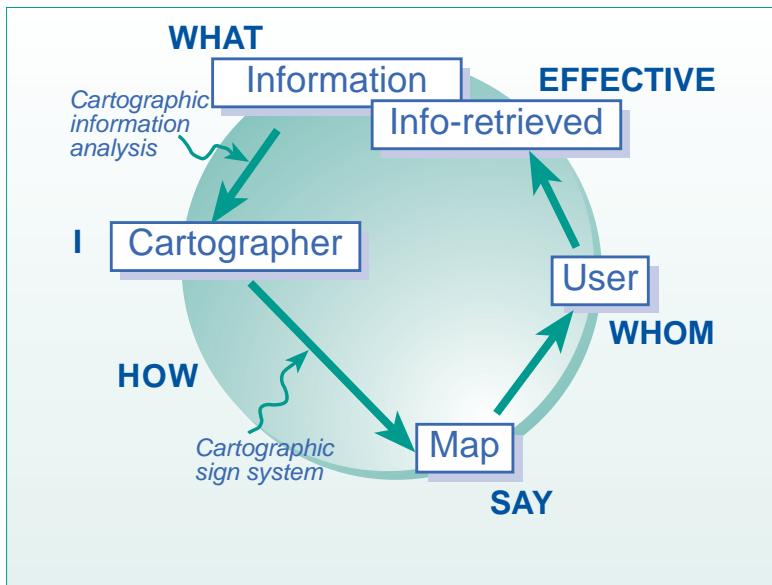


Figure 7.10: The cartographic communication process, based on “How do I say what to whom, and is it effective?”
Source: Figure 5–5 in [30].

Returning to the scheme, the map (the medium that does 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 ‘Info-retrieved’. From the figure it becomes clear that the boxes with ‘Information’ and ‘Info-retrieved’ do not overlap. This means the information derived by the map user is not the same as the information that the cartographic communication process started with. There may be several causes. Possibly, the original information was not all used or additional information has been added during the process. Omission of information could be deliberately caused by the cartographer, with the aim of emphasizing the remaining information. Another possibility is that the map user did not fully understand the map. 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 them to combine this prior knowledge with the knowledge retrieved from the map.

7.4 The cartographic toolbox

7.4.1 What kind of data do I have?

To derive 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 land use, collected in 2005, the title could be *Landuse of ... 2005*. Secondly, the individual component(s), such as landuse, and probably relief, should be analysed and their nature described. Later, these components should be visible in the map legend.

Cartographic data analysis

We have already discussed different kinds of data values on page [75](#), in relation to the types of computations we can do on them. Here we take a look at the different types of data in relation to how we might map or display them.

Data will be of a *qualitative* or *quantitative* nature. Qualitative data is also called *nominal* or *categorical* data. This data exists as 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 represented as basic geographic units: 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 is 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 does have a known absolute zero. An example is income: someone earning \$100 earns twice as much as someone with an income of \$50. In order to generate maps, quantitative data are often classified into categories according to some mathematical method.

In between qualitative and quantitative data, one can distinguish *ordinal data*. These data are measured along a relative scale, 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 7.1.

<i>Measurement scale</i>	<i>Nature of data</i>
Nominal, categorical	Data of different nature / identity of things (qualitative)
Ordinal	Data with a clear element of order, though not quantitatively determined (ordered)
Interval	Quantitative information with arbitrary zero
Ratio	Quantitative data with absolute zero

Table 7.1: Differences in the nature of data and their measurement scales

7.4.2 How can I map my data?

Basic elements of a map, irrespective of the medium on which it is displayed, are point symbols, line symbols, area symbols, and text. The appearance of point, line, and area symbols can vary depending on their nature. Most maps in this book show symbols in different size, shape and colour. Points can vary in form or colour to represent the location of shops or they can vary in size to represent aggregated values (like number of inhabitants) for an administrative area. Lines can vary in colour to distinguish between administrative boundaries and rivers, or vary in shape to show the difference between railroads and roads. Areas follow the same principles: difference in colour distinguishes between different vegetation stands.

Symbology

Although the variations in symbol appearance are only limited by the imagination they can be grouped together in a few categories. Bertin [5] distinguished six categories, which he called the *visual variables* and which may be applied to point, line and area symbols. As illustrated in Figure 7.11, they are:

- *Size,*
 - *Value (lightness),*
 - *Texture,*
 - *Colour,*
 - *Orientation and*
 - *Shape.*
- Visual variables

differences in:	symbols		
	point	line	area
size			
value			
grain			
colour			
orientation			
shape			

Figure 7.11: Bertin's six visual variables illustrated.
Source: Plate 1 in [31].

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 influence the map user's perception in different ways. What is perceived depends on the human capacity to see or perceive:

- What is of equal importance (e.g. all red symbols represent danger),
- Order (e.g. the population density varies from low to high—represented by light and dark colour tints, respectively),
- Quantities (e.g. symbols changing in size with small symbols for small amounts), or
- An instant overview of the mapped theme.

There is an obvious relationship between the nature of the data to be mapped and the 'perception properties' of visual variables. In Table 7.2, the measurement scales as defined in Table 7.1 are linked to the visual variables displayed in Figure 7.11. 'Dimensions of the plane' is added to the list of visual variables; it is the basis, used for the proper location of symbols on the plane (map). The perception properties of the remaining visual variables have been added. The next section discusses some typical mapping problems and demonstrates the above.

perception properties	visual variables	measurement scales			
		nominal	ordinal	interval	ratio
	dimensions of the plane	x	x	x	x
order & quantities	size		x	x	x
order	(grey) value		x	x	
	grain/textture		x	x	
equal importance	colour hue	x			
	orientation	x			
	shape	x			

Table 7.2: Measurement scales linked to visual variables based on perception properties

7.5 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 nevertheless commonly found.

7.5.1 How to map qualitative data

If, after a long fieldwork period, one 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 7.12: A good example of mapping qualitative data

The application of colour would be the best solution since it 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 7.12 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 much higher, the map, at the scale displayed here, will become too cluttered. The map

Readability

can also be made by filling the watershed areas by different forms (like small circles, squares, triangles, etc.) in one colour (e.g. black for a monochrome map)—as an application of the visual variable shape. The amount of geographic units that can be displayed is then even more critical.

Figure 7.13 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 ‘value’. Looking at the map may cause perceptual confusion since the map image suggests differences in importance that are not there in reality. In Figure 7.13(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.

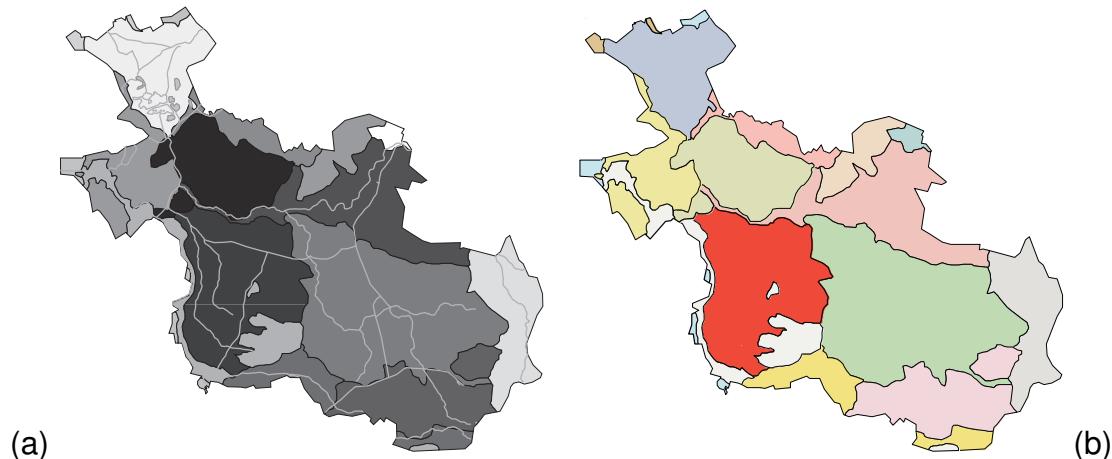


Figure 7.13: Two examples of wrongly designed qualitative maps: (a) misuse of tints of black; (b) misuse of bright colours

7.5.2 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 7.14 shows the final map for the province of Overijssel.

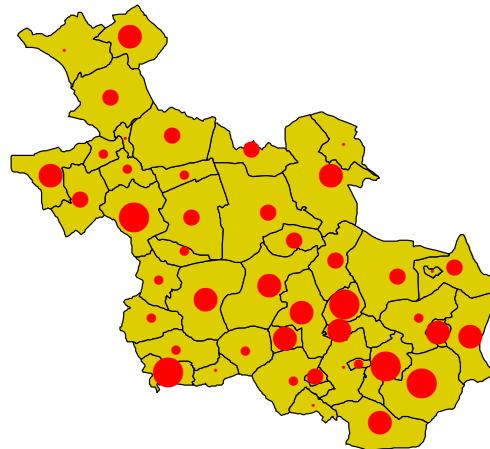


Figure 7.14: Mapping absolute quantitative data

The fact that it is easy to make errors can be seen in Figure 7.15. In 7.15(a), different tints of green (the visual variable ‘value’) have been used to represent *absolute* population numbers. The reader might get a reasonable impression of the individual amounts but not of the actual geographic distribution of the population,

as the size of the geographic units will influence the perceptual properties too much. Imagine a small and a large unit having the same number of inhabitants. The large unit would visually attract more attention, giving the impression there are more people than in the small unit. Another issue is that the population is not necessarily homogeneously distributed within the geographic units. Colour has also been misused in Figure 7.15(b). The applied four-colour scheme makes it impossible to infer whether red represents more populated areas than blue. It is impossible to instantaneously answer a question like “Where do most people in Overijssel live?”

Misuse of colour

On the basis of absolute population numbers per municipality and their geographic size, we can also generate a map that shows population density per municipality. We then deal with *relative quantitative data*. The numbers now have

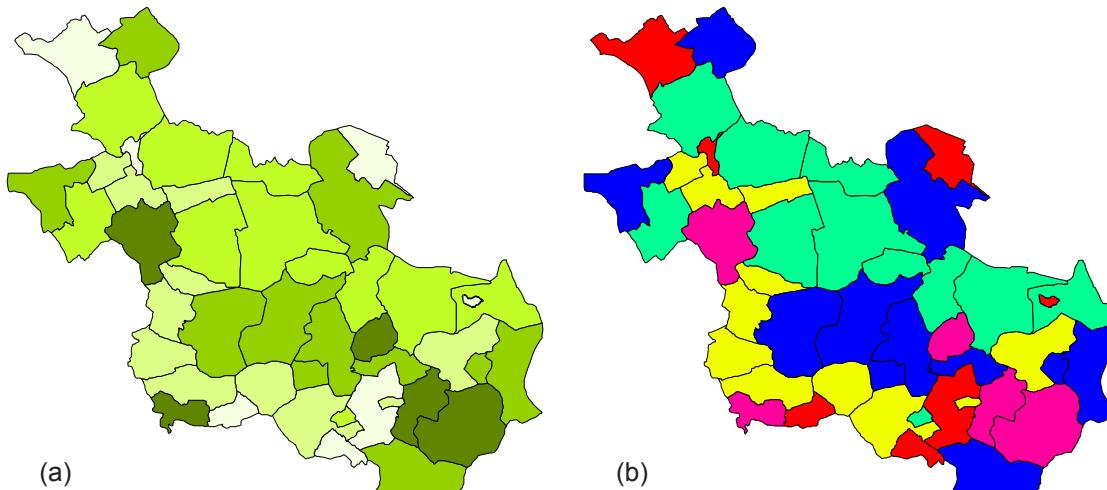


Figure 7.15: Poorly designed maps displaying absolute quantitative data: (a) wrong use of green tints for absolute population figures; (b) incorrect use of colour

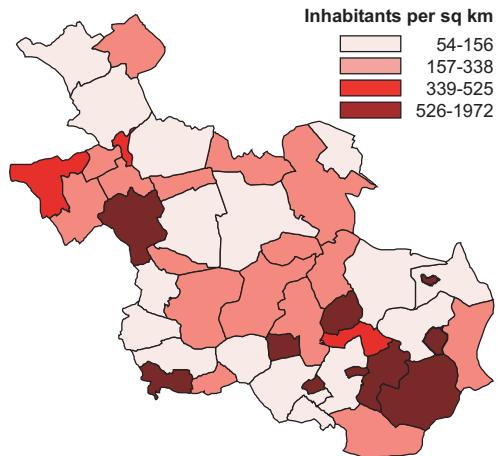


Figure 7.16: Mapping relative quantitative data

a clear relation with the area they represent. The geographic unit will again be municipality. The aim of the map is to give an overview of the distribution of the population density. In the map of Figure 7.16, 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 7.17(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 than dark-red?

In Figure 7.17(b) colour has been used in combination with value. The first impression of the map reader would be to think the brown areas represent the areas

Mapping relative quantities

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 studies the badly designed maps carefully, the information can 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).

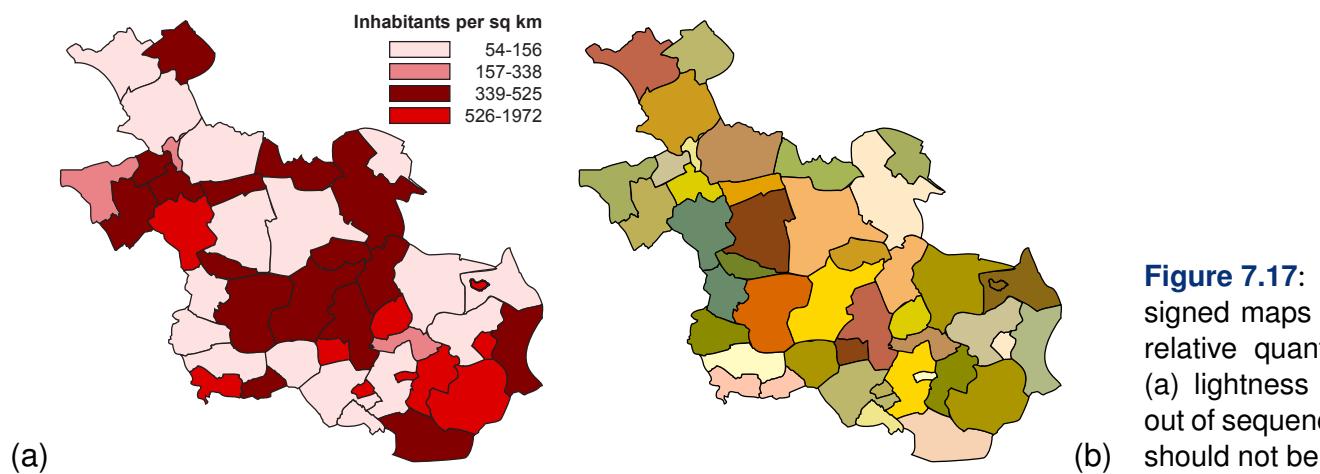


Figure 7.17: Badly designed maps representing relative quantitative data: (a) lightness values used out of sequence; (b) colour should not be used

7.5.3 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 the space between the contour lines can be filled with colour and value information following a convention, e.g. green for low elevation and brown for high elevation areas. This technique is known as hypsometric or layer tinting. Even more advanced is the addition of *shaded relief*. This will improve the impression of the three-dimensional relief (see Figure 7.18).

The shaded relief map uses the full three-dimensional information to create shading effects. This map, represented on a two-dimensional surface, can also be floated in three-dimensional space to give it a real three-dimensional appearance of a 'virtual world', as shown in Figure 7.18(d). Looking at such a representation one can immediately imagine that it will not always be effective. Certain (low) objects in the map will easily disappear behind other (higher) objects.

Interactive functions are required to manipulate the map in three-dimensional space in order to look behind some objects. These manipulations include panning, zooming, rotating and scaling. Scaling is needed, particularly along the *z*-axis, since some maps require small-scale elevation resolution, while others require large-scale resolution, i.e. vertical exaggeration. One can even imagine that other geographic, three-dimensional objects (for instance, the built-up area of a city and individual houses) have been placed on top of the terrain model,

Three dimensional
appearance

Manipulating 3D maps

like it is done in Google Earth. Of course, one can also visualize objects below the surface in a similar way, but this is more difficult because the data to describe underground objects are sparsely available.

Socio-economic data can also be viewed in three dimensions. This may result in dramatic images, which will be long remembered by the map user. Figure 7.19 shows the absolute population figures of Overijssel in three dimensions. Instead of a proportionally sized circles to depict the number of people living in a municipality (as we did in Figure 7.14) the proportional height of a municipality now indicates total population. The image clearly shows that Enschede (the large column in the lower right) is by far the highest populated municipality.

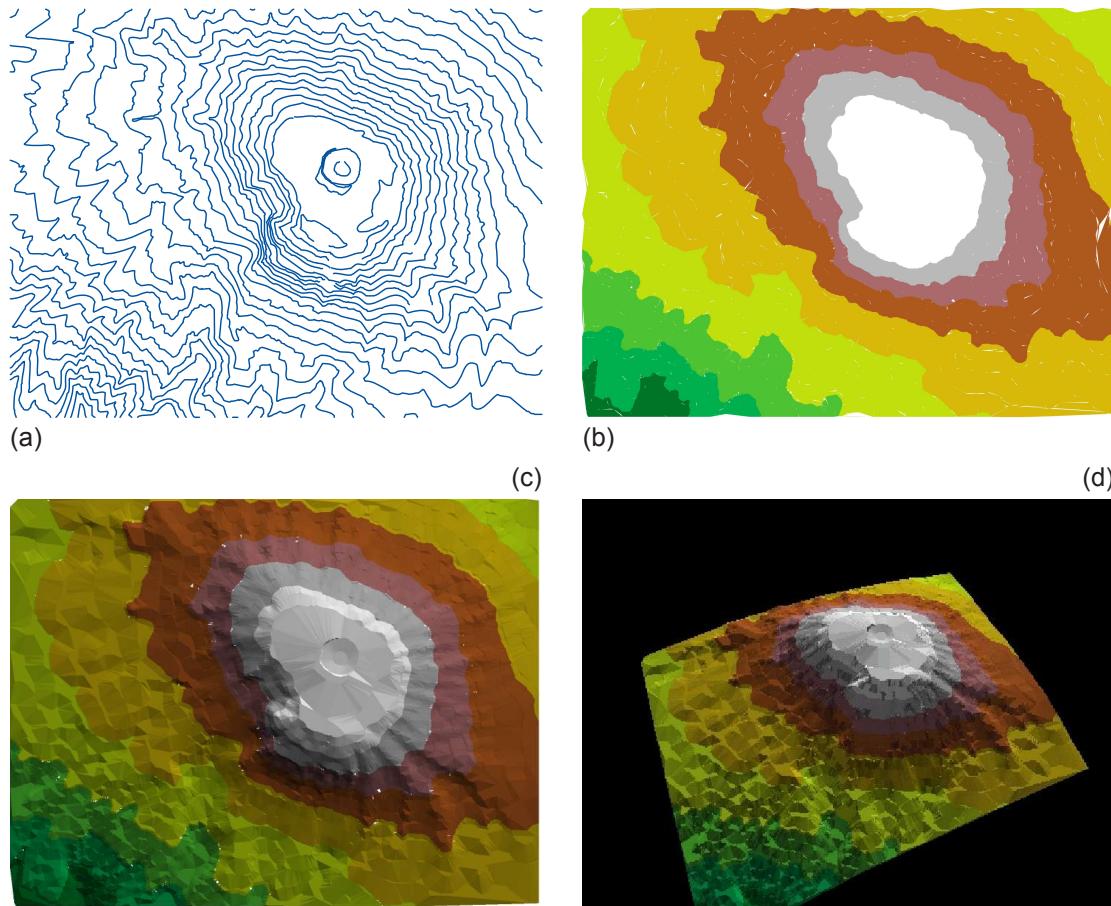


Figure 7.18: visualization of terrain elevation: (a) contour map; (b) map with layer tints; (c) shaded relief map; (d) 3D view of the terrain

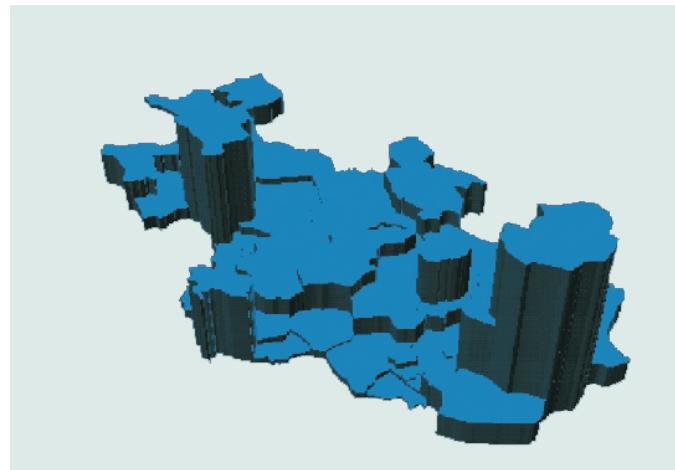


Figure 7.19: Quantitative data visualized in three dimensions

7.5.4 How to map time series

Advances in spatial data handling have not only made the third dimension part of GIS routines. Nowadays, the handling 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 only static paper maps.

Mapping time means mapping change. This may be change in a feature's geometry, in its attributes or both. Examples of changing geometry are the evolving coastline of the Netherlands (as displayed in Figure 7.3), the location of Europe's national boundaries, or the position of weather fronts. The changes of a land parcel's owner, landuse, 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 are 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. Size changes can also be applied to other point and line symbols to show increase and decrease over time. Specific point symbols such as 'crossed swords' (battle) or 'lightning' (riots) can be found to represent dynamics in historic maps. Another alternative is the use of the visual variable value (expressed as tints). In a map showing the development of a town, dark tints represent old built-up areas, while new built-up areas are represented by light tints (see

Mapping changing phenomena

Using symbology to represent change

Figure 7.20(a)).

It is possible to distinguish between three temporal cartographic techniques (see Figure 7.20):

1. *Single static map*: Specific graphic variables and symbols are used to indicate change or represent an event. Figure 7.20(a) applies the visual variable value to represent the age of the built-up areas;
2. *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 should be limited since it is difficult for the human eye to follow long series of maps (Figure 7.20(b));
3. *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 can be deduced from real ‘change’ in the image itself, not from a spatial sequence (Figure 7.20(c)).

Temporal cartographic
techniques

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?”

User interaction

Most of the general software packages for viewing animations already offer facilities such as 'pause' (to look at a particular frame) and '(fast-)forward' and '(fast-)backward', or step-by-step display. More options have to be added, such as a possibility to directly go to a certain frame based on a task like: 'Go to 1850'.

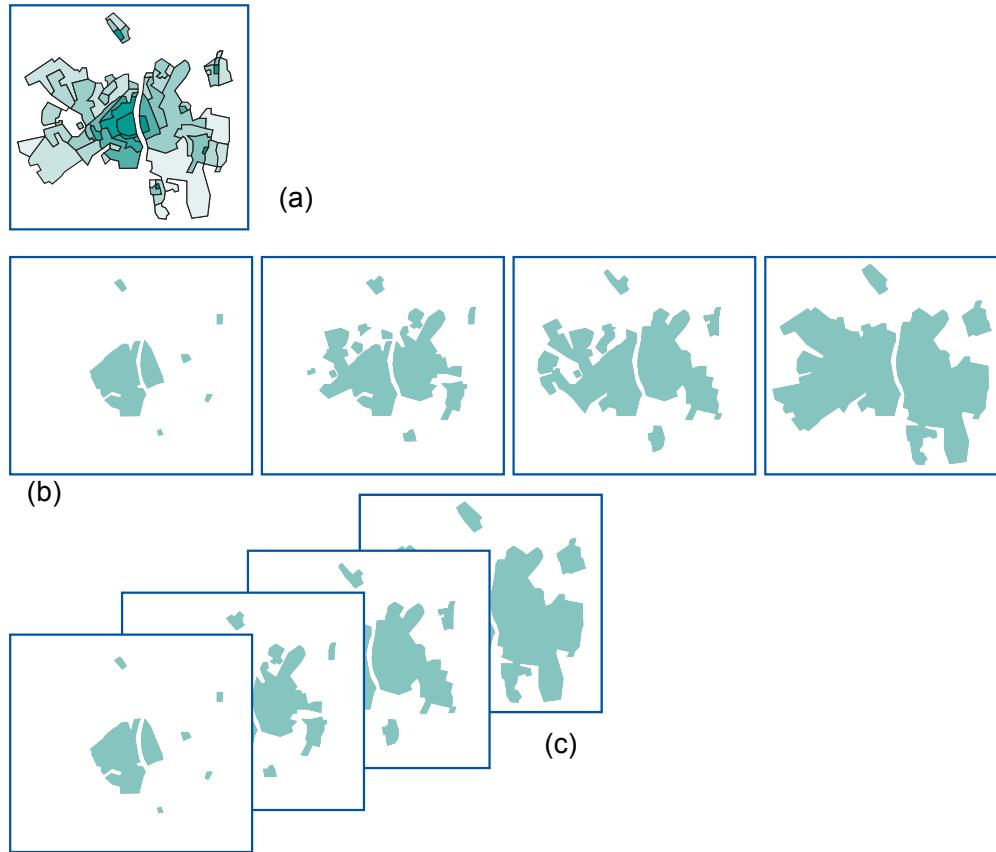


Figure 7.20: Mapping change; example of the urban growth of the city of Maastricht, The Netherlands: (a) single map, in which tints represent age of the built-up area; (b) series of maps; (c) (simulation of an) animation.

7.6 Map cosmetics

Most maps in this chapter are correct from a cartographic grammar perspective. However, many of them lack the additional information needed to be fully understood that is usually placed in the margin of printed maps. Each map should have, next to the map image, a *title*, 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 datum* and *map projection* used, and some *lineage* information, (such as data sources, dates of data collection, methods used, etc.).

Fundamental requirements

Further information can be added that indicates when the map was issued, and by whom (author / publisher). All this information allows the user to obtain an impression of the quality of the map, and is comparable with metadata describing the contents of a database or data layer.

Figure 7.21 illustrates these map elements. On paper maps, these elements (if all relevant) 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 titles are often available on demand as well.

Space constraints

The map in Figure 7.21 is one of the first in this chapter that has text included. Figure 7.22 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 7.22 an example can be found. Italics—*cf.* the visual variable of orientation—have been used for building names to distinguish them

Text

from road names. Another common example is the use of colour to differentiate (at nominal level) between hydrographic names (in blue) and other names (in black). The text should also be placed in a proper position with respect to the object to which it refers.

Maps constructed via the basic cartographic guidelines are not necessarily visually appealing maps. Although well-constructed, they might still look sterile. The design aspect of creating appealing maps also has to be included in the visualization process. ‘Appealing’ does not only mean having nice colours. One of the keywords here is *contrast*. Contrast will increase the 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 concept. The need for visual hierarchy in a map is best understood when looking at the map in Figure 7.23(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.

Contrast and visual hierarchy

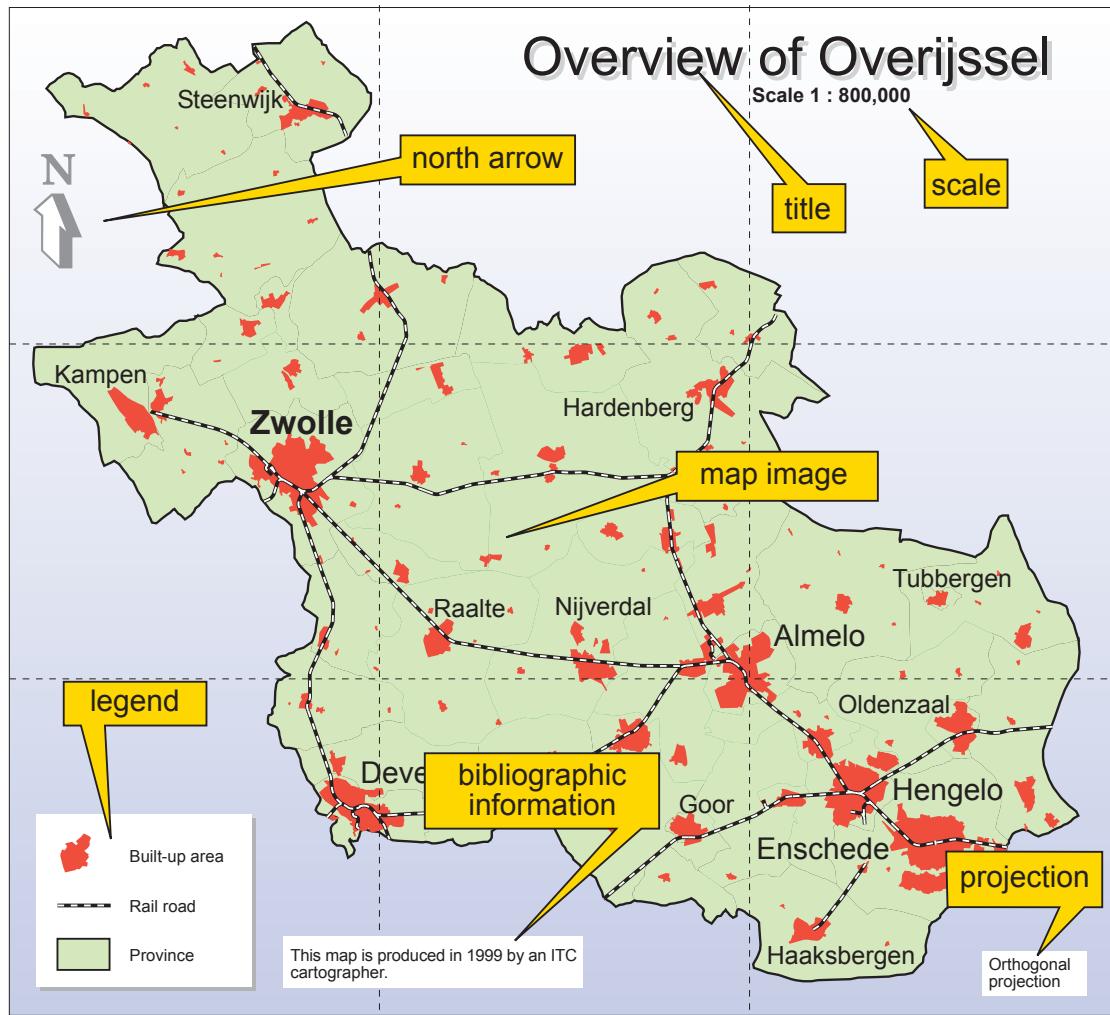


Figure 7.21: The paper map and its (marginal) information. Source: Figure 5–10 in [30].

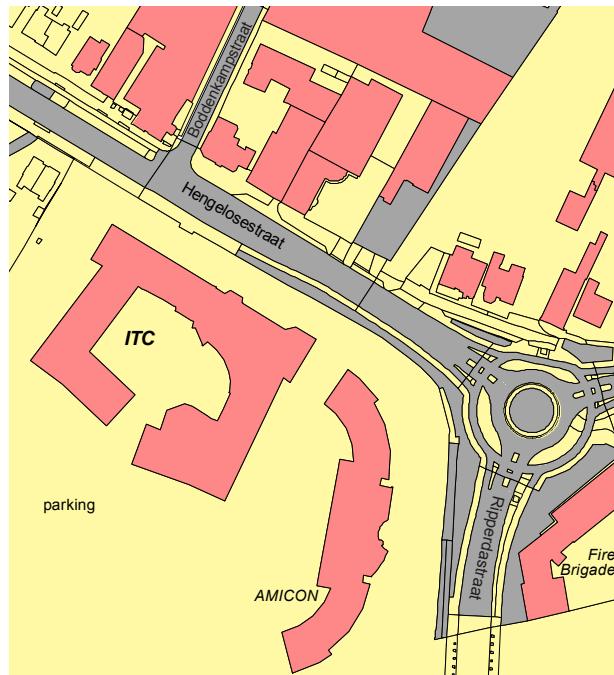


Figure 7.22: Text in the map



Figure 7.23: Visual hierarchy and the location of the ITC building: (a) hierarchy not applied; (b) hierarchy applied

7.7 Map dissemination

The map design will not only be influenced by the nature of the data to be mapped or the intended audience (the ‘what’ and ‘whom’ from “How do I say What to Whom, and is it Effective”), the output medium also plays a role. Traditionally, maps were produced on paper, and many still are.

Currently, 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 their 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 also be made available.

On-screen maps

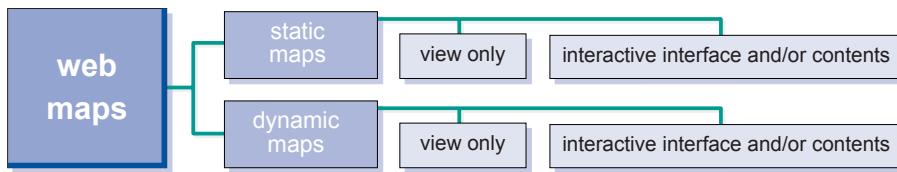
Maps and multimedia (photography, sound, video, animation) can be integrated. Some of today’s electronic atlases, such as the *Encarta World Atlas* are good examples of how multimedia elements can be integrated with the map. Pointing to a country on a world map starts the national anthem of the country 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.

Multimedia maps

The World Wide Web is nowadays a common medium used to present and disseminate spatial data. Here, maps can play their traditional role, for instance to show the location of objects, or provide insight into spatial patterns, but because of the nature of the internet, 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. Maps can also be used as ‘previews’ of spatial data products to be acquired through a spatial data clearinghouse that is part of a Spatial Data Infrastructure. For that purpose we can make use of geo-webservices which can provide interactive map views as intermediate between data and web browser (please refer to Section 3.2.3).

Maps as visual interfaces



See also kartoweb.itc.nl/webcartography/webmaps/classification.htm

Figure 7.24:
Classification of maps
on the WWW. Source:
Figure 1–2 in [30].

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 7.24) 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. Many *static maps* on the web are view-only. Organizations, such as map libraries or tourist information providers, often make their maps available in this way. This form of presentation can be very useful, for instance, to make historical maps more widely accessible. 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.

Static maps

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.

Dynamic maps

Slightly more interactive versions of this type of map are those to be played by media players, for instance those in *QuickTime* format, or as a Flash movie. Plug-ins to the web 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), for instance, can be used for this purpose. It stores a true 3D model of the objects, not just a series of 3D views.

Interactive maps

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 has discussed 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 anima-

tions. Animations are well suited to display spatial change.

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 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 7.8 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 482, we discussed three techniques for mapping changes over time. We already discussed the issue of change detection, and illustrated it in Figure 2.25. 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.

Bibliography

- [1] ABBOTT, E. A. *Flatland—A romance of many dimensions*. Penguin Group, New York, N.Y., 1984.
- [2] ARBIA, G., GRIFFITH, D., AND HAINING, R. Error propagation modelling in raster gis: overlay operations. *International Journal of Geographical Information Science* 12, 2 (1998), 145–167. [435](#)
- [3] ARONOFF, S. *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa, Canada, 1989. [32](#), [344](#)
- [4] BELWARD, A. S., AND VALENZUELA, C. R., Eds. *Remote Sensing and Geographical Information Systems for Resource Management in Developing Countries*. Kluwer Academic, Dordrecht, The Netherlands, 1991.
- [5] BERTIN, J. *Sémiology Graphique*. Mouton, Den Haag, The Netherlands, 1967. [460](#), [466](#)

- [6] BIJKER, W. *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. [128](#), [130](#)
- [7] BOARD, C. Report of the working group on cartographic definitions. *Cartographic Journal* 29, 1 (1990), 65–69. [446](#), [520](#)
- [8] BONHAM-CARTER, G. F. *Geographic information systems for geoscientists : Modeling with GIS*, vol. 13 of *Computer methods in the geosciences*. Pergamon, Kidlington, U.K., 1994. [377](#)
- [9] BURROUGH, P. A. *Principles of Geographical Information Systems for Land Resources Assessment*. Monographs on Soil and Resources Survey. Clarendon Press, Oxford, U.K., 1986.
- [10] BURROUGH, P. A. Natural objects with indeterminate boundaries. In *Geographic objects with indeterminate boundaries*, P. A. Burrough and A. U. Frank, Eds. Taylor and Francis, London, U.K., 1996, pp. 3–28. [295](#)
- [11] BURROUGH, P. A., AND McDONNELL, R. *Principles of Geographical Information Systems*. Oxford University Press, Oxford, U.K., 1998. [337](#)
- [12] CARTWRIGHT, W., PETERSON, M., AND GARTNER, G., Eds. *Multimedia Cartography*. Springer, Berlin, Germany, 1999.
- [13] CHRISMAN, N. R. Errors in categorical maps: testing versus simulation. In *Proceedings AutoCarto* (1989), pp. 521–529. [434](#)
- [14] CLARKE, D. G., AND CLARK, M. Lineage. In *Elements of Spatial Data Quality*, S. C. Guptill and J. L. Morrison, Eds. Elsevier Science, Oxford, U.K., 1995, pp. 13–30. [301](#)

- [15] DATE, C. J. *An Introduction to Database Systems*, seventh ed. Addison-Wesley, Reading, Ma, 2000.
- [16] DENT, B. D. *Cartography: Thematic Map Design*, fifth ed. WCB/McGraw-Hill, Boston, Ma, 1999.
- [17] DEVILLERS, R., AND JEEANSOULIN, R. *Data Structures and Algorithms*. ISTE Ltd, London, United Kingdom, 2006. [284](#)
- [18] DiBIASE, D. Visualization in earth sciences. *Earth and Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences* 59, 2 (1990), 13–18. [457](#)
- [19] ELMASRI, R., AND NAVATHE, S. B. *Fundamentals of Database Systems*, second ed. Benjamin/Cummings, Redwood City, Ca, 1994.
- [20] HEARNSHAW, H. M., AND UNWIN, D. J., Eds. *Visualization in Geographical Information Systems*. John Wiley & Sons, London, U.K., 1994.
- [21] HEUVELINK, G. B. M. *Error propagation in quantitative spatial modelling—Applications in Geographical Information Systems*. Nederlandse Geografische Studies. Koninklijk Aardrijkskundig Genootschap, Utrecht, 1993. [297, 435](#)
- [22] HOULDING, S. W. *3D Geoscience Modeling: Computer Techniques for Geological Characterization*. Springer-Verlag, Berlin, Germany, 1994.
- [23] HUNTER, G. J., AND BEARD, K. Understanding error in spatial databases. *The Australian Surveyor* 37, 2 (1992), 108–119. [433](#)
- [24] ILIFFE, J. *Datums and Map Projections for Remote Sensing, GIS and Surveying*. Whittles Publishing, CRC Press, 2000. [226](#)

- [25] ILWIS DEPARTMENT. *ILWIS 2.1 for Windows—User's Guide*. ITC, Enschede, The Netherlands, 1997.
- [26] KAINZ, W. Logical consistency. In *Elements of Spatial Data Quality*, S. C. Guptill and J. L. Morrison, Eds. Elsevier Science, Oxford, U.K., 1995, pp. 109–137.
- [27] KIIVERI, H. T. Assessing, representing and transmitting positional accuracy in maps. *International Journal of Geographical Information Systems* 11, 1 (1997), 33–52. [435](#)
- [28] KNIPPERS, R. A., AND HENDRIKSE, J. Coordinaattransformaties. *Kartografisch Tijdschrift* 3 (2000). [233](#)
- [29] KRAAK, M.-J. Exploratory cartography, maps as tools for discovery. *ITC Journal 1998*, 1 (1998), 46–54.
- [30] KRAAK, M.-J., AND BROWN, A., Eds. *Web cartography, developments and prospects*. Taylor & Francis, London, U.K., 2000. [445](#), [452](#), [458](#), [461](#), [487](#), [491](#)
- [31] KRAAK, M.-J., AND ORMELING, F. J. *Cartography: Visualization of Spatial Data*. Addison-Wesley Longman, London, U.K., 1996. [467](#)
- [32] KRAAK, M.-J., AND ORMELING, F. J. *Cartography: Visualization of Spatial Data*, second ed. Pearson Education, Harlow, U.K., 2003. [447](#), [456](#)
- [33] LANGRAN, G. *Time in Geographic Information Systems*. Technical Issues in Geographic Information Systems. Taylor & Francis, London, U.K., 1992. [127](#)
- [34] LAURINI, R., AND THOMPSON, D. *Fundamentals of Spatial Information Systems*, vol. 37 of *The APIC Series*. Academic Press, London, U.K., 1992. [152](#)

- [35] LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. M., AND RHIND, D. W., Eds. *Geographical Information Systems: Principles, Techniques, Management, and Applications*, second ed., vol. 1. John Wiley & Sons, New York, N.Y., 1999. [456](#)
- [36] MACEACHREN, A. M., AND TAYLOR, D. R. F., Eds. *Visualization in Modern Cartography*. Pergamon Press, London, U.K., 1994.
- [37] MCCORMICK, B., DEFANTI, T. A., AND (EDS.), M. D. B. Visualization in scientific computing. *ACM SIGGRAPH Computer Graphics—Special issue 21*, 6 (1987). [457](#)
- [38] MEIJERINK, A. M. J., DE BROUWER, J. A. M., MANNAERTS, C. M., AND VALENZUELA, C. R. *Introduction to the Use of Geographic Information Systems for Practical Hydrology*, vol. 23 of *ITC Publication*. ITC, Enschede, The Netherlands, 1994.
- [39] MOLENAAR, M. *An Introduction to the Theory of Spatial Object Modelling*. Taylor & Francis, London, U.K., 1998.
- [40] MORRISON, J. L. Topographic mapping for the twenty-first century. In *Framework of the World*, D. Rhind, Ed. Geoinformation International, Cambridge, U.K., 1997, pp. 14–27. [458](#)
- [41] NATIONAL MAPPING DIVISION, U. S. GEOLOGICAL SURVEY. Spatial data transfer standard. Tech. rep., U. S. Department of the Interior, 1990. [288](#)
- [42] NEBERT, D., Ed. *Developing Spatial Data Infrastructures: The SDI Cookbook v2.0*. Global Spatial Data Infrastructure (GSDI),

- <http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf>, 2004.
146
- [43] OPENSHAW, S., CHARLTON, M., AND CARVER, S. Error propagation: a Monte Carlo simulation. In *Handling geographical information: methodology and potential applications*, I. Masser and M. Blakemore, Eds. Longman, Harlow, U.K., 1991, pp. 78–101. 294
- [44] PEUQUET, D. J., AND MARBLE, D. F., Eds. *Introductory Readings in Geographic Information Systems*. Taylor & Francis, London, U.K., 1990.
- [45] PFLUG, R., AND HARBAUGH, J. W., Eds. *Three-dimensional Computer Graphics in Modeling Geologic Structures and Simulating Geologic Processes*, vol. 41 of *Lecture Notes in Earth Sciences*. Springer-Verlag, Berlin, Germany, 1992.
- [46] PREPARATA, F. P., AND SHAMOS, M. I. *Computational Geometry—An Introduction*. Springer-Verlag, New York, NY, 1985. 94
- [47] RAPER, J., Ed. *Three dimensional Applications in Geographic Information Systems*. Taylor & Francis, London, U.K., 1989.
- [48] SAMET, H. *Applications of Spatial Data Structures*. Addison-Wesley, Reading, Ma, 1990.
- [49] SAMET, H. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, Ma, 1990. 153
- [50] SLOCUM, T.A., M. R. B., AND KESSLER, F. *Thematic Cartography and Geovisualization*, third ed. Pearson Education, USA, 2009. 456

- [51] SNIJDER, J. P. Map projections - a working manual. Professional paper 1395, U.S. Geological Survey, 1987. [219](#)
- [52] STAR, J., AND ESTES, J. *Geographic Information Systems, An Introduction.* Prentice Hall, Englewood Cliffs, NJ, 1990.
- [53] TEMPFLI, K., Ed. *Principles of Remote Sensing*, fourth ed., vol. 2 of *ITC Educational Textbook Series*. International Institute for Geo-Information Science and Earth Observation, Enschede, The Netherlands, 2008. [34](#), [63](#), [121](#), [227](#), [272](#), [299](#), [309](#)
- [54] TOMLIN, C. D. *Geographic Information Systems and Cartographic Modeling.* Prentice Hall, Englewood Cliffs, NJ, 1990. [381](#)
- [55] TURNER, A. K., Ed. *Three-dimensional Modeling with Geoscientific Information Systems.* Kluwer Academic, Dordrecht, The Netherlands, 1992.
- [56] VEREGIN, H. Developing and testing of an error propagation model for GIS overlay operations. *International Journal of Geographical Information Systems* 9, 6 (1995), 595–619. [434](#), [435](#)
- [57] WORBOYS, M. F. *GIS: A Computing Perspective.* Taylor & Francis, London, U.K., 1995.
- [58] ZADEH, L. A. Fuzzy sets. *Information and Control* 8 (1965), 338–353. [296](#)
- [59] ZEILER, M. *Modeling our World—The ESRI Guide to Geodatabase Design.* ESRI Press, Redlands, Ca, 1999.

Glossary

[previous](#)[next](#)[back](#)[exit](#)[contents](#)[index](#)[glossary](#)[web links](#)[bibliography](#)[about](#)

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) .
- 2½D** 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.
- ArcInfo** A GIS software package developed in the 1980's and 1990's 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.
- CIO** Conventional International Origin. The mean position of the pole in the year 1903 (based on observations between 1900 and 1905) used to compensate for changes in position of the Earth's rotational axis over time (referred to as polar motion).
- DBMS** Database Management System.
- DEM** Digital Elevation Model.
- dpi** Dots per inch; the unit of scanner (or printer) resolution, expressed as how many pixels can be read (printed) per inch.
- DTM** Digital Terrain Model.
- e.g.** For example, ; (*exempli gratia*).

- ESRI** Environmental Systems Research Institute, Inc. The American company, based in Redlands, California, that created and develops ArcGIS.
- GDOP** Geometric Dilution Of Precision.
- GIS** Geographic Information System.
- GLONASS** Global Navigation Satellite System. The American satellite-based positioning system.
- GMT** Greenwich Mean Time.
- GPS** Global Positioning System.
- GRS80** Geodetic Reference System 1980.
- HSDPA** High-Speed Downlink Packet Access: A protocol for (fast) cellular phone data transmission.
- i.e.** That is, ; meaning, ; (*id est*).
- ILWIS** Integrated Land and Water Information System. A GIS software package developed in the 1980's and 1990's at ITC. Historically more raster-based than vector-based.
- ISO** International Organization for Standardization.
- ITRF** International Terrestrial Reference Frame.
- ITRS** International Terrestrial Reference System.

- NDVI** Normalized Difference Vegetation Index.
- NOAA** National Oceanic Atmospheric Administration; an institute falling under the U.S. Department of Commerce which monitors the Earth's environment through satellite imagery.
- OGC** Open Geospatial Consortium.
- SA** Selective Availability.
- SDSS** Spatial Decision Support System(s).
- 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.
- TAI** International Atomic Time.
- TIN** Triangulated Irregular Network.
- UT** Universal Time.
- UTC** Coordinated Universal Time.
- viz.** Namely, ; (*videlicet*).
- WGS84** World Geodetic System 1984.

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

Agent-Based Model (ABM) These attempt to model processes in the form of multiple (possibly interacting) agents (which might represent individuals) using sets of decision-rules about what the agent can and cannot do. As such, a key notion is that simple behavioral rules for individual agents generate complex behaviour for the entire 'system'. Agent-based models have been developed to understand aspects of complex systems, for example by incorporating *stochastic* and/or *deterministic* components.

Algorithm A procedure used to solve a mathematical or computational problem, or to address a data processing issue. Algorithms usually consist of a set of rules written in a computer language.

Altitude The elevation of an object above a reference surface, usually mean sea level.

Aspect The geographical direction toward which a slope faces, measured in degrees from north, in a clockwise direction.

Attribute Data associated with a spatial feature or sample location, stored as a column in a database table. The name of the column should suggest what the values in that column stand for. These values are known as *attribute values*.

Autocorrelation see 'spatial autocorrelation'.

Azimuth In mapping and navigation, azimuth is the direction to a target with respect to north and usually expressed in degrees.

Bandwidth The range of frequencies in a radio signal. The wider, the more data can be carried.

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.

Categorical data See 'Nominal data'.

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.

Channel In satellite-based positioning, the circuitry of the receiver that allows to receive the signal of a single satellite.

Check point An additional ground point used to independently verify the degree of accuracy of a geometric transformation (e.g. georeferencing, aerial triangulation).

Clearinghouse Centralized repository, often forming part of a Spatial Data Infrastructure, where data users can 'shop' for spatial data. Clearinghouses usually have an entrance through the world wide web referred to as a 'web portal'.

Clock bias In satellite-based positioning, the difference between a receiver's clock reading and that of the (largely synchronized) satellite clock(s).

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 line An elevation isoline. Valuable especially in map production.

Contour map Map in which contour lines are used to represent terrain elevation.

Control segment The worldwide network of satellite monitoring and control stations that ensure accuracy of satellite positions and clocks.

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.

Deterministic (In the context of an *application model*), a procedure or function that generates an outcome with no allowance or consideration for variation. Deterministic models are good for predicting results when the input is predictable, and the exact functioning of the 'process' is known. The opposite of *stochastic*.

Digital Elevation Model A representation of a surface in terms of elevation values that change with position. Elevation can refer to the ground surface, a soil layer, etc. According to the original definition data should be in a raster format.

Digital Terrain Model (DTM). A digital representation of terrain relief in terms of (x, y, z) coordinates and possibly additional information (on break-lines and salient points). Usually, z stands for elevation, and (x, y) for the horizontal position of a point. To the concept of a DTM it does not

matter whether z is orthometric or ellipsoidal elevation. Horizontal position can be defined by geographic coordinates or by grid coordinates in a map projection. DTM data can be given in different forms (contour lines, raster, TIN, profiles, etc).

Dilution of precision A factor of multiplication that (negatively) affects the ranging error in satellite-based positioning. It is caused by a non-optimal geometry of the satellites used for positioning in the receiver.

Doppler aiding A technique that satellite positioning receivers use to improve their reception of satellite signals, and to improve the accuracy with which they determine velocity of the receiver, on the basis of a measured *Doppler effect*.

Doppler effect The change in frequency of a radio signal caused by the relative motion of the transmitter with respect to the receiver.

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.

Error matrix The matrix that compares samples taken from the data to be evaluated with observations that are considered as correct (reference). The error matrix allows calculation of quality parameters such as overall accuracy, error of omission, and error of commission.

Euclidean space A space in which locations are identified by coordinates, and with which usually the standard, Pythagorean *distance* function between locations is associated. Other functions, such as *direction* and *angle*, can also be present. Euclidean space is n -dimensional, and we must make a choice of n , being 1, 2, 3 or more. The case $n = 2$ gives us the *Euclidean plane*, which is the most common Euclidean space in GIS use.

Evapotranspiration (Sometimes erroneously written as evapotransporation); the process by which surface water, soils, and plants release water vapour to the atmosphere through evaporation (surface water, soils) 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.

Filter In the context of this book, an algorithm for eliminating, reducing, attenuating or extracting information from raster data; see also filtering.

Filtering Computational process of changing given values such that a contained component is either eliminated, reduced, attenuated, or extracted. Examples include extracting slope information from an elevation raster using x - and y -gradient filters, or extracting boundaries from polygons represented as rasters.

Foundation data see 'base data'.

Geo-webservices Software programs that act as an intermediate between web users and geographic data(bases). These can vary from a simple map display service to a service which involves complex spatial calculations.

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 (also: 'Geoinformation'). 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 entry, management, analysis 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, was or will be; an abbreviation of 'geographically referenced'.

Geospatial data Data related to locations on (the surface of) the Earth. 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 phenomena 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.

Global Positioning System The American satellite-based positioning system. More generally, satellite surveying method providing accurate geodetic coordinates for any point on the Earth at any time.

Granularity The level of detail with which something is represented.

Grid A network of regularly spaced horizontal and perpendicular lines (as for locating points on a map). We may associate (field) values with the nodes of the grid. 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.

Ground Control Point (GCP). A ground point reliably identifiable in the image(s) under consideration. It has known coordinates in a map or terrain coordinate system, expressed in the units (eg, meters, feet) of the specified coordinate system. GCPs are used for georeferencing and image orientation.

Image In the context of this book, this term refers to raw data produced by an electronic sensor, which are not pictorial, but arrays of digital numbers related to some property of an object or scene, such as the amount of reflected light. An image may comprise any number of bands. When the reflectance values have been translated into some ‘thematic’ variable we refer to it as a raster. An image consists of pixels, whereas a raster is composed of cells.

Integer Any ‘whole’ number in the set $\{\dots, -2, -1, 0, 1, 2, \dots\}$; computers cannot represent arbitrarily large numbers, and some maximum (and minimum) integer is usually indicated.

Interpolation (From Latin *interpolare*, putting in between). Estimating the value of a (continuous) variable that is given by n sampled values at some intermediate point or instant. See ‘spatial interpolation’.

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.

Kernel In the context of this book, a *window* of a given size and shape used in ‘moving-window’ operations on point data, or a *neighbourhood* of n by m cells used in operations on raster data. See ‘filter’.

Latitude/Longitude The coordinate components of a spherical coordinate system, referred to as *geographic coordinates*. The latitude is zero on the equator and increases towards the two poles to a maximum absolute value of 90° . The longitude is counted from the Greenwich meridian positively eastwards to the maximum of 180° .

Least-squares adjustment A method of correcting observations in which the sum of the squares of all the residuals derived by fitting the observations to a mathematical model is minimised. Least squares adjustment is based on probability theory and requires a (large) number of redundant measurements.

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.

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 A simplified, purpose-specific graphical representation of geographic phenomena, usually on a planar display. Defined in [7] as “A tool for presenting geographic information in a way that is visual, digital or tactile.”.

Map coordinate system A system of expressing the position of a point on the Earth’s surface by planar rectangular coordinates using a particular map projection, such as UTM, the Lambert’s conical projection, or an azimuthal stereographic projection (as used in the Netherlands).

Map generalization The meaningful reduction of map content to accommodate scale decrease.

Map projection The functional mapping of a curved horizontal reference surface onto a flat 2D plane, using mathematical equations.

Map scale The ratio of distance on the map to the corresponding horizontal distance in ‘real world’ units. The ratio is commonly expressed as $1 : m$, where m is the scale factor (e.g. 1:25,000).

Multi-path error A *ranging error* that occurs when the satellite signal is received multiple times and these receptions interfere. This is normally caused by reflections off objects.

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 when the values are sorted according to some set of non-overlapping categories.

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.

Orbit The path followed by one body (e.g. a satellite) in its revolution about another (e.g. the Earth).

Ordinal data Data values that serve to identify or name something, and for which some natural ordering of the values exists. No arithmetic is possible on these data values.

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).

Pseudorange In satellite-based positioning, a distance measurement obtained by a receiver from a satellite's signal. Uncorrected for *clock bias*.

Ranging error In satellite-based positioning, the error made when a receiver determines the distance to a satellite.

Raster A set of regularly spaced (and contiguous) cells with associated (field) values. In contrast to a *grid*, the associated values represent *cell* values, not point values. This means that the value for a cell is assumed to be valid for all locations within the cell. 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.

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.

Relative positioning (Also: differential positioning) Determination of position using another receiver with accurately known position that is tracking the same satellite signals.

Sampling Selecting a representative part of a population for statistical analysis; to this end various strategies can be applied, such as random sampling, systematic sampling, stratified sampling, etc.

Satellite A manufactured vehicle intended to orbit the earth, or another celestial body.

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.

Space segment The constellation of satellites that can be used for positioning.

Spatial autocorrelation The principle that locations which are closer together are more likely to have similar values than locations that are far apart. Often referred to as *Tobler's first law of Geography*.

Spatial data In the broad sense, any data with which position is associated. See *geospatial data*.

Spatial Data Infrastructure (SDI); The relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data.

Spatial 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 allows users to store, query and manipulate collections of georeferenced data.

Spatial interpolation Any technique that allows to infer some unknown property value of a spatial phenomenon from values for the same property of nearby spatial phenomena. The underlying principle is that nearby things are most likely rather similar. Many spatial interpolation techniques exist.

Spatial Reference Frame (SRF): A physical realization 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 (SRS): 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.

Stochastic (In the context of an *application model*), a random or probability-based model component that generates different results from some initial value, depending on the probability function over time. The opposite of deterministic approaches or models. Used when we do not know the exact functioning of a process.

String Any sequence of characters chosen from the alphabet plus a set of other characters like punctuation 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 represents *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 Topology refers to the spatial relationships between geographical elements in a data set that do not change under a continuous transformation.

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.

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. Pronounced as 'tapl'.

User segment The community of users and their satellite receivers, in satellite-based positioing.

Visual variable (Also: graphic variable); an elementary way in which graphic symbols are distinguished from each other. Commonly, the following six visual variables are recognized: *size, (lightness) value, texture, colour, orientation and shape*.

Web portal World wide web based entrance to a spatial data clearinghouse.

Index

accumulated flow count raster, 393
accuracy, 41, 275, 277–290, 302, 303
 attribute, 288
 location, 279
 positional, 277
 temporal, 290
animated map, 472
application model, 40, 414
area object, 88
area size, 341, 344
attribute, 76, 114, 119, 132, 143, 152,
 154, 156, 288, 298, 348–351
autocorrelation
 spatial, 73
base data, 304
boundary, 88, 89, 94, 97
 crisp, 71
 fuzzy, 71
boundary model, 89

buffer, 336
buffer zone, 336, 385, 387, 423, 427
cartographic generalization, 334
cartographic grammar, 437, 458
cartography, 430–482
categorical data, 65
cellular automata, 416
centroid, 77, 341
change detection, 118
classification, 358–364
 automatic, 363
 equal frequency, 363
 equal interval, 363
 user-controlled, 361
classification operator, 334
classification parameter, 358
clearinghouse, 271
conformal map projection, 213
connectivity, 87, 112, 337, 405–413

consistency
 temporal, 290
 topological, 99
contour line, 107, 337
contour map, 467
control point, 265
control segment, 227
coordinate systems
 planar, 196
 spatial, 196
coordinate thinning, 309
Coordinated Universal Time, 234
coupling
 embedded, 418
 loose, 418
 tight, 418

data
 3D, 101
 georeferenced, 35
 geospatial, 35
 spatial, 35, 72–115
 spatiotemporal, 116–121
 thematic, 85
data layer, 88, 114, 312, 335, 361, 366, 386, 421, 437
data preparation, 294
data quality, 35, 274, 421

data standards, 36, 273
database, 39, 43
 geo-, 45
 spatial, 45–175
datum
 global, 192
 local, 190
datum transformation, 221–225
deductive approaches, 417
Delaunay triangulation, 388
diffusion computation, 384, 392
diffusion function, 390
digitizing, 265–269
 automatic, 265
 manual, 265
 semi-automatic, 265
dilution of precision, 242
dimension
 geographic, 21
 spatial, 21
 temporal, 21
dissolve, 295
distance, 344
dynamic map, 471–473

edge matching, 307
ellipsoid, 186
embedded coupling, 418

equidistant map projection, 213
equivalent map projection, 213
error, 106
 propagation, 419, 420
Euclidean plane, 57

facet, 102

field
 continuous, 60, 62, 77, 82, 105, 141,
 310, 312, 315, 316
 differentiable, 62
 discrete, 60, 62, 141, 310, 312, 313
 geographic, 59, 62–63

filter, 399

filtering, 399

flow computation, 384, 393

flow direction raster, 393

Galileo, 253

generalization
 cartographic, 436

geo-webservices, 137, 271, 274

geodatabase, 168, 172, 175

geographic information science, *see*
 GIS, *see* GIS

geographic information system, *see* GIS

Geoid, 183

geoinformatics, 33

geoinformation, 72
geometric transformation, 308
georeferenced, 29, 33, 35, 57, 58
geostatistics, 327
GIS, 16, 21–23, 33–34, 130
 definition of, 22
GLONASS, 252
GPS, 249–251
grid, 76

height, 186–467
 orthometric, 184
hillshading, 61, 396
horizontal datum, 186
hypsometric tinting, 467

image, 262
inductive approaches, 417

information
 geographic, 16, 33, 35–37
interior, 94, 97
International Terrestrial Reference Frame,
 193, 226
International Terrestrial Reference System, 192, 254
interpolation, 28, 29, 49, 72, 77, 84,
 106, 310, 311, 315, 321, 327,
 336

- IDW, 322
- trend surface, 316
- interval data, 65, 454
- inverse distance weighting, 322
- isoline, 32, 107, 312, 320, 321
- join condition, 163
- kernel, 399
- kriging, 327
- large-scale, 103
- latitude, 197
- length
 - of polyline, 341, 344
- levelling
 - geodetic, 184
- line object, 86
- line segment, 86
- lineage, 291
- local resistance raster, 390
- location, 341, 344
 - object, 67
- longitude, 197
- loose coupling, 418
- map, 431–438, 456–482
 - large-scale, 435
 - small-scale, 435
- thematic, 437
- topographic, 436, 437
- map algebra, 371
- map generalization, 444
- map grid, 202
- map legend, 475
- map output, 480–482
- map projection, 207–308
 - changing, 219
- map scale, 41, 103, 435
- map theme
 - physical, 437, 440
 - socio-economic, 437, 440
- map title, 475
- mapping equation
 - forward, 208
 - inverse, 209
- mean sea level, 183
- measurement, 339–344
- metadata, 36, 272, 291, 475
- metric, 94
- minimal bounding box, 342
- minimal cost path, 391
- model, 53, 333
 - agent-based, 416
 - aggregate, 416
 - application, 414

dynamic, 416
individual, 416
process, 416
static, 416
model generalization, 334
modelling, 39, 53, 414
moving window averaging, 322
multi-path reception, 240
multi-representation spatial data, 305
multi-scale spatial data, 304

NDVI, 374
neighbourhood function, 336, 382–393
network allocation, 410–411, 413
network analysis, 337, 405–413
network direction, 405
network function, 337
network partitioning, 406, 410
network trace analysis, 411–413
Niña, La, 20, 31
Niña, La, 19
Niño, El, 19–21, 24, 25, 31, 38, 40, 43,
 44, 49
nominal data, 65, 454
normal map projection, 212

object
 geographic, 60, 67–70, 109

oblique map projection, 212
optimal path finding, 407–408
ordinal data, 65, 455
orientation
 object, 67
overlay function, 335, 366–380
 on raster data, 371–380
 on vector data, 367–369
overshoot, 295

phenomenon
 dynamic, 116
 geographic, 19, 21, 45, 57, 59–70
pixel, 263
point object, 85
polygon clipping operator, 368
polygon intersection, 367
polygon overwrite operator, 368
polyhedron, 102
positional fix, 229
positioning
 2D and 3D, 231
 absolute, 228
 network, 246
 relative, 244
 satellite-based, 226–256
precision, 275
primary data, 262

proximity function, 384–388
pseudorange, 228, 237, 239, 240
Pythagorean distance, 341

quadtree, 78, 105
qualitative data, 454
quantitative data, 454, 464
query, 154

raster, 75, 105, 262
raster calculus, 371
raster cell, 263
raster resolution, 325
rasterization, 299
ratio data, 66, 454
reclassification, 358
redundancy
 data, 89
reference surface, 182
regression, 316
relation, 152, 154, 156
relational data model, 154–164
relationship
 topological, 97
resolution, 29
retrieval operator, 334
root mean square error, 279

SDI, 136, 271, 481

SDSS, 145
search window, 336
secondary data, 264
selected object, 346
selection object, 346
selective availability, 237
shape
 object, 67
simplex, 96
simplicial complex, 96, 99
size
 object, 67
sliver polygon, 304
slope, 336
slope angle, 396, 401
slope aspect, 396, 402, 403
slope convexity, 396
slope gradient, 401
small-scale, 103
solid, 101
space, 57
 Euclidean, 57, 94
 geographic, 16, 17, 45, 49
 metric, 94
 topological, 94
space segment, 227
spatial aggregation, 359

- spatial analysis, 46
- spatial autocorrelation, 73, 79, 323, 327
- spatial data, 17
- spatial data infrastructure, 136
- spatial dissolving, 359
- spatial information theory, 132
- spatial join, 368
- spatial merging, 359
- spatial reference system, 29
- spatial selection, 345–357
 - interactive, 346
 - using distance, 353
 - using topology, 352
- spatio-temporal, 21
- standards, 136
- static map, 472, 481
- surface
 - secant, 210
- tangent surface, 210
- tessellation, 74–79
 - irregular, 78
 - regular, 75
- Thiessen polygon, 313, 388
- tie point, 265
- tight coupling, 418
- time
 - concepts of, 117
- representing in GIS, 117
- TIN, 82, 106
- tolerance, 283
- topological mapping, 93
- topology, 174, 301
 - spatial, 91–99
- transformations
 - coordinate, 217
- transverse map projection, 212
- trend surface, 320
- triangulation, 83, 321
 - Delaunay, 84
- trilateration, 229
- tuple, 152, 154, 156
- turning cost table, 407
- undershoot, 295
- user segment, 227
- vector, 74, 109
- vectorization, 265, 267, 299, 312
- vertical datum, 184
- visibility function, 337
- visual hierarchy, 476
- visual interpretation, 266
- visual variable, 456
- visualization, 430–482
 - web portal, 271

WWW, 480

x-gradient filter, 402

y-gradient filter, 402

Appendix A

Internet sites

General GIS sites

- The Open GIS Consortium (OGC) homepage
- GIS dot com, ESRI site
- Institut Géographique National, France
- United States Geological Survey (USGS)
- Harvard University list of GIS sites

Spatial data sources (maps and data)

- The European INSPIRE web portal
- The Dutch National Atlas online
- National Geographic's Maps and Geography pages
- Digital Chart of the World, at Pennsylvania State University, U.S.A.
- Pennsylvania State University Libraries, Maps Library
- Ordnance Survey, United Kingdom
- United States "Geospatial One Stop" web portal
- United States Geological Survey (USGS) National Geologic Map Database
- ESRI's Data Repository
- Worldwide National Statistical Offices
- University of Texas at Austin, On-line Map Gallery
- Refdesk dot com on Atlases and Maps

Spatial reference systems and frames

- Geometric Aspects of Mapping; Division of Cartography, ITC
- Active GPS Reference System for the Netherlands (AGRS.NL)
- ITRF homepage
- SAtelliten POsitionierung System (SAPOS)
- GeodIS page, maintained by Deutsches Geodätisches Forschungsinstitut (DGFI), Geodetic Reference System 1980 (GRS80)
- International Earth Rotation and Reference Systems service
- Office of the Surveyor General of Land Information New Zealand (LINZ), guide to datums, projections and heights
- Ordnance Survey of Great Britain. A Guide to Coordinate Systems in Great Britain

Other useful links and guides

- The Open GIS Consortium “Learning Resources” page
- ColorBrewer, a useful online guide to using colour in maps and graphics
- FreeGIS.org - free GIS software and data
- The Generic Mapping Tools site
- The Geographer’s Craft GIS notes
- NCGIA Core Curriculum in GIScience